Pharmacophylogenomics

Explaining interspecies differences in drug discovery

Een wetenschappelijke proeve op het gebied van de
Natuurwetenschappen, Wiskunde en Informatica

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op vrijdag 14 september 2007
om 15.30 uur precies
door

Tim Hulsen

geboren op 10 augustus 1979
te Wijchen

Promotor:

Prof. dr. Jacob de Vlieg

Copromotor:

Dr. Peter M.A. Groenen (NV Organon)

Manuscriptcommissie:

1. Prof. dr. Han G. Brunner
2. Prof. dr. Antoine H.C. van Kampen (AMC, UvA, NBIC)
3. Prof. dr. Peter van der Spek (Erasmus MC)

ISBN/EAN 978-90-9022031-4

© 2007 T. Hulsen

All rights reserved. Save exceptions stated by the law, no part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, included a complete or partial transcription, without the prior written permission of the author, application for which should be addressed to author.

Alle rechten voorbehouden. Niets uit deze opgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand of openbaar worden gemaakt, in enige vorm of op enige wijze, zonder voorafgaande schriftelijke toestemming van de auteur.

Front cover:

“Gene Therapy” © Mehau Kulyk, ANP / Science Photo Library

Back cover:

Waalkade, Nijmegen (location of the PhD party) © Kees Hulsen

Invitation:

Kruittoren, Kronenburgerpark, Nijmegen (near to the author’s home) © Kees Hulsen

Printed by:

Drukkerij Print Partners Ipskamp te Enschede

Table of contents

Chapter 1

General introduction

1.1 Introduction

1.1.1 Goal

The current drug discovery pipeline can be regarded as slow and inefficient. The average time spent, from the very start of the process until the final clinical trials, is around fourteen years [1]. The 'attrition rate', i.e. 1 - the ratio approved drugs / tested compounds [2], is currently ~ 0.9873: one marketable drug emerges from approximately eighty screened compounds (figure 1 of [3]). This high attrition rate is mainly caused by the low success rates in the first clinical phases. In this introduction, we build an hypothesis on how this pipeline can be shortened by the application of new genomics technologies in the drug discovery process. The first paragraphs will discuss genomics in general, its appraisal and acceptance, including important techniques such as sequence comparison and ortholog identification, whereas the latter paragraphs attempt to enforce the assumption that genomics applications can improve the drug discovery pipeline in a more pragmatic way.

1.1.2 Genomics

In the past 30 years, starting with the sequencing of bacteriophage Φ-X174 in 1977 [4], the genomes from almost 300 organisms have been fully sequenced [5]. This wealth of information facilitates genome-wide analyses that have been impossible to do before. This relatively new field within biology is generally referred to as 'genomics'. This term is derived from the word 'genome', which is a contraction created by Hans Winkler (1920) of the words 'gene' and 'chromosome'. As defined by Wikipedia [6]: "Genomics is the study of an organism's genome and the use of the genes. It deals with the systematic use of genome information, associated with other data, to provide answers in biology, medicine, and industry." A whole range of –omics fields exists: e.g. transcriptomics (dealing with RNA), proteomics (dealing with proteins), metabolomics (dealing with metabolic pathways). This approach of connecting large biological data sets is also referred to as the 'systems biology' approach, in which biology is not seen as a number of loose components, but more as a complex system. Figure 1 shows that the use of the words 'genomics', 'transcriptomics', 'proteomics', 'metabolomics' and 'systems biology' in the PubMed [7] literature database has increased drastically over the past ten years, reflecting the growing use of these techniques. Especially the popularity of the word 'genomics' has been growing almost exponentially: from 0 in 1986 to 4339 (0.64% of the total number of articles) in 2005. However, this is probably just the start, since more and more genomics data is becoming available. Figure 1 also shows that the use of the word 'pharmacogenomics' does not increase as much as the other five words, displaying that the effects of genomics have been mostly limited to fundamental science until now. Despite this observation, genomics is expected to gain influence on applied fields such as pharmaceutics [8] and medicine [9] in the future.

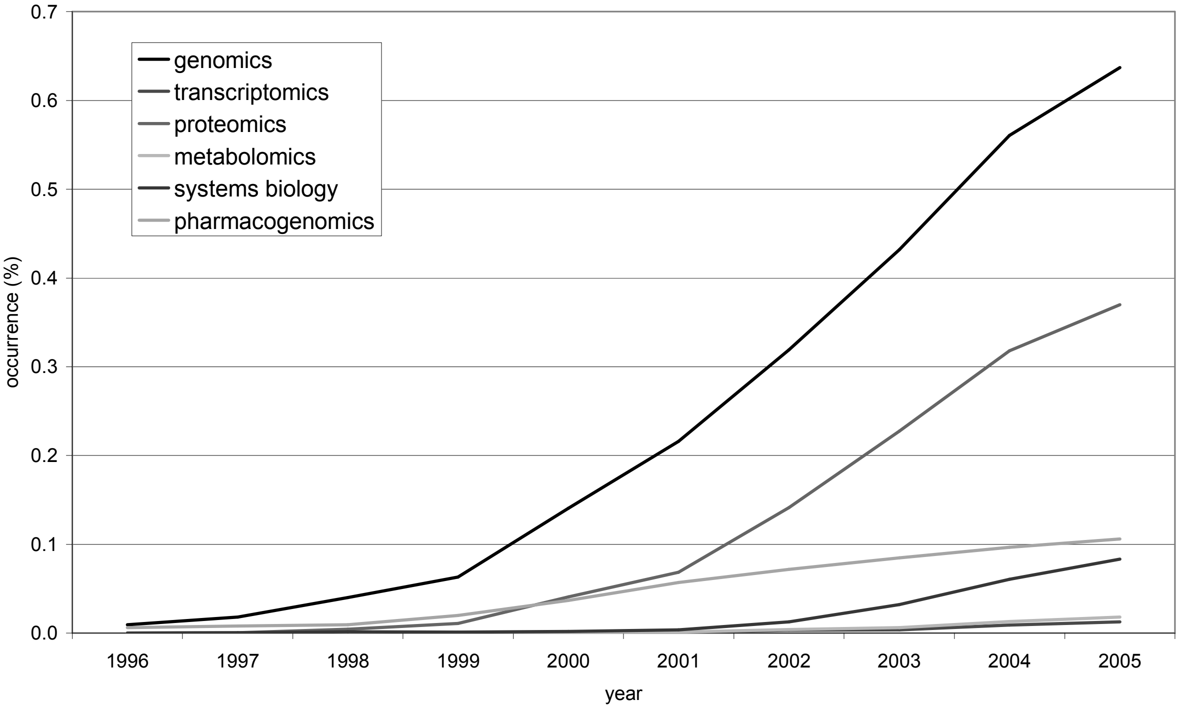


Figure 1. Popularity of -omics search terms in the PubMed database
The percentage of articles (titles + abstracts) in the PubMed database that contain the words ‘genomics’ (black line), ‘transcriptomics’ (purple line), ‘proteomics’ (blue line), ‘metabolomics’ (green line), ‘systems biology’ (red line) or ‘pharmacogenomics’ (orange line). Horizontal axis: year. Vertical axis: number of articles that contain that specific search term, divided by the total number of articles published in that year (in %). *Color version on page 146.*

1.1.3 Comparative genomics

Genomics opens doors to new ways of research and novel methodologies, such as the usage of all-against-all sequence comparisons [10]. These all-against-all sequence comparisons can be applied for cross-species analyses, which are summarized by the term ‘comparative genomics’: the study of relationships between the genomes of different species. Examples of comparative genomics methodologies include pairwise and multiple sequence alignments, phylogenetic trees, phylogenetic patterns [11] and gene order conservation [12]. The main goal of comparative genomics is to gain a better understanding of how species have evolved and to determine the function of genes and non-coding regions of the genome of a certain species, by using the information from the genomes of other species.

Comparative biology, and subsequently comparative genomics, originated in comparative embryological studies performed in the nineteenth century, mainly by Ernst Haeckel. According to these studies different vertebrates show strikingly similar developmental stages, regardless of the taxon concerned. Although the credibility of Haeckel’s work on embryological evolution has been undermined by recent findings [13], the main conclusions from his work are still valid, and his (and other) work on embryology and evolution has had large consequences. Extrapolation of experimental evidence from model-organisms allowed a detailed understanding of human embryology. Together with natural occurring mutants with congenital defects, many developmental pathways in different species have been discovered. Since thirty years, it has been possible to not only study the embryological phenotypes, but also the underlying molecular mechanisms. This new field is called evolutionary developmental biology, or 'evo-devo' [14].

The major principles of comparative genomics are straightforward [15]. Common features of multiple organisms will often be encoded within the DNA that is conserved between the species. More accurately, the DNA sequences encoding the RNA sequences and proteins responsible for functions that were conserved from the last common ancestor should be preserved in contemporary genome sequences. Likewise, the DNA sequences controlling the expression of genes regulated similarly in two related species should be conserved as well [16]. Conversely, regulatory elements responsible for interspecies differences will themselves be divergent.

Due to the large amounts of data analyzed in the field of comparative genomics, the application of computers is essential. This is why the terms ‘computational genomics’ and ‘comparative genomics’ are sometimes interchanged erroneously. Computational genomics refers simply to genomics studies that are carried out in silico completely, while comparative genomics refers to a much wider range of genomics studies, carried out over multiple species.

1.2 Orthology

1.2.1 Introduction

Homology is a very important concept in evolutionary biology, phylogenomics and comparative genomics. The concept arose within comparative morphological and paleontological systematics around 150 years ago and was later on used in evolutionary biology as synapomorphic similarity inherited from a common ancestor [17]. In genomics, it refers to two or more genes or proteins that share a common ancestor. Homology is usually measured by the rate of similarity at the sequence level. Although two very similar sequences do not necessarily need to be homologs, and two proteins that are very dissimilar can be homologs, this sequence similarity is still considered to be a quite reliable measurement [18].

Homology can be subdivided into orthology, paralogy and xenology. The term orthology describes the evolutionary relationship between homologous genes whose independent evolution reflects a speciation event, whereas paralogy refers to genes that have diverged from a common ancestor through a gene duplication event [19]. Orthology is often misused as a description of functionally equivalent genes in different species. Orthologs do not necessarily have the same function, although they are very likely to have a functional similarity. Orthologous genes are more likely to have a functional similarity than paralogous genes, which have often undergone changes in substrate or ligand specificity [20, 21]. Figure 2 explains these concepts through the evolution of the globin gene. The third subdivision of homology, xenology, refers to homology that arises via horizontal (or lateral) gene transfer between unrelated species [22].

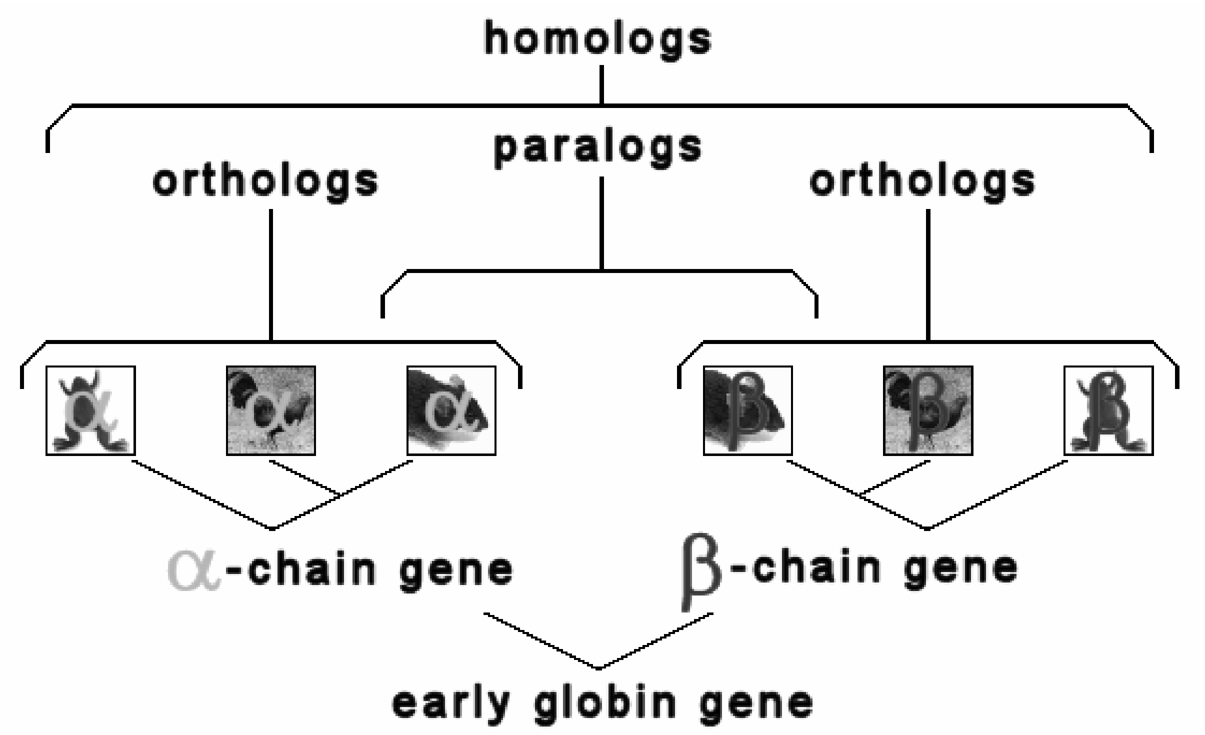


Figure 2. The concepts of homology, orthology and paralogy explained by the example of the globin gene
Orthologs and Paralogs are two types of homologous sequences. Orthology describes genes in different species that derive from a common ancestor. Orthologous genes may or may not have the same function. Paralogy describes homologous genes within a single species that diverged by gene duplication.

1.2.2 Functional annotation

The high level of functional conservation between orthologous proteins makes orthology highly relevant for protein function prediction. It is also widely used in genome analysis, where the information about a protein in one species is used for the functional annotation of the orthologous protein in another species. Figure 3 shows that currently about 42.4% of the human genes have an unknown molecular function. By defining orthologous relationships between genes in model organisms that have a known molecular function and these human genes, we are able to transfer the function to the human gene with a high degree of certainty. This is known as 'homology-based function prediction' [23]. Of course, there is a large chance that both genes in an orthologous pair have an unknown function. By creating orthologous groups, in which multiple species are included, this chance can be decreased: in many cases, at least one of the genes in one of the species has a function assigned, which makes it possible to assign this function to all of the genes in the orthologous group [24]. Homology-based function prediction should be regarded as a supplement to experimental annotation, since it is less reliable [25, 26] and needs at least one of the genes or proteins in an orthologous group to have an experimental annotation.

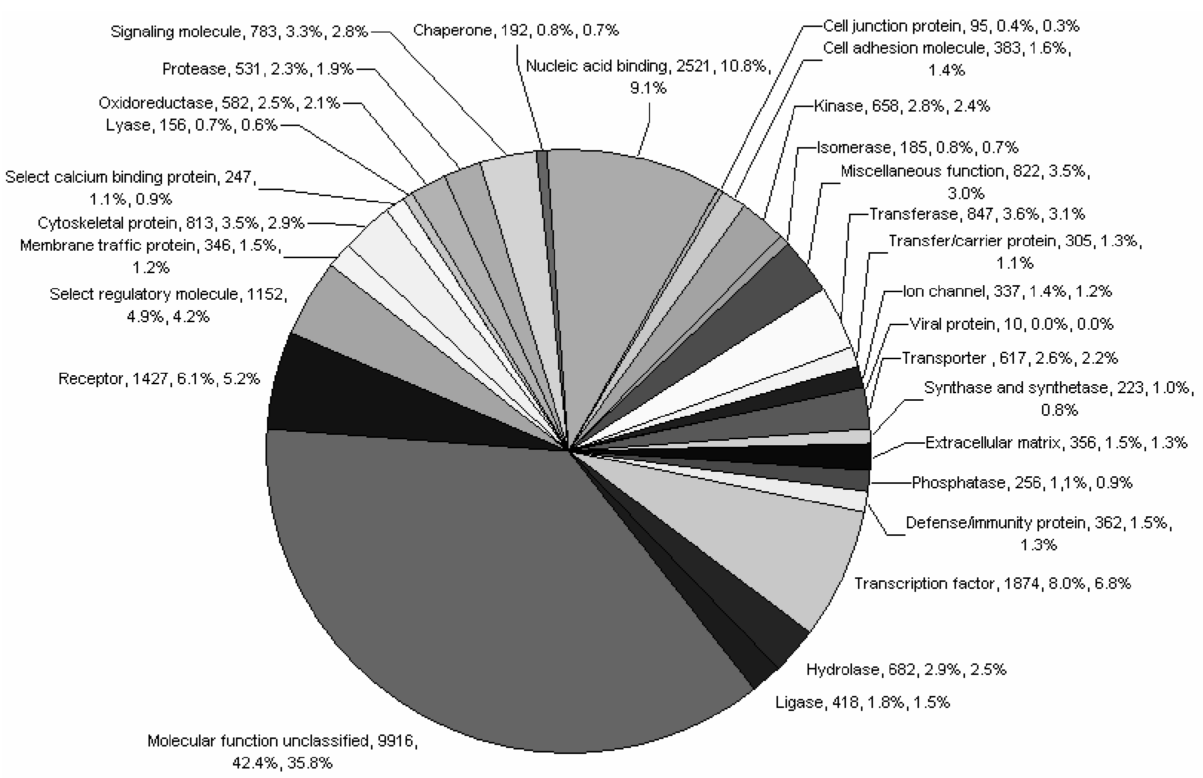


Figure 3. Distribution of the 27,686 molecular functions of 23,401 human genes

Each slice lists the number of human gene functions assigned to a given category of molecular function (in brackets: percentages of total number of molecular functions and of total number of genes). The categories are provided by the Panther classification from Applied Biosystems [27].

1.2.3 Ortholog identification methods

Several methods have been developed for finding orthologous relationships. Most methods rely on sequence identity or similarity. The easiest method is the ‘best bidirectional hit’ or ‘best reciprocal hit’ method. This method needs as input a list of all genes in genome A compared to all genes in genome B, and vice versa. The best hit for gene A1 from genome A in genome B is detected: gene B1. If the comparison of gene B1 with genome A returns the same gene A1, these two genes are considered to be best directional hits and thus orthologs. There are other ortholog identification methods that are extensions to this approach, like InParanoid [28] and its successor MultiParanoid [29]. These methods do not only find the best bidirectional hits, but also check for paralogs within the genomes and co-orthologs between the genomes. This resembles the reality of evolution much better, because it allows for one-to-many and many-to-many orthologous relationships. Other methods such as COG [30] and OrthoMCL [31] use cluster algorithms to create orthologous groups. Some algorithms do not only rely on pairwise sequence comparison but also on an extra step consisting of multiple sequence alignment and/or phylogenetic trees. Examples of this are the Phylogenetically Inferred Groups (PhIGs) [32] and COCO-CL [33]. Finally, there are methods that use synteny information for detecting orthologous pairs. Genes are often conserved in clusters over the genome, which means that the genomic context of a gene can help in finding orthologs. The Homologene [34] method is an example of an algorithm in which sequence comparison is used in combination with synteny information.

1.3 Sequence comparison

1.3.1 Introduction

As discussed in the previous section, sequence comparison constitutes the basis for most of the ortholog identification methods. It enables researchers to find proteins with a high similarity, which are highly probable to be homologous [18]. The field of sequence comparison originated in the development of protein-sequencing methods in the 1950s [35] and the assembly of protein sequence databases in the 1970s (PIR, [36]). These protein sequence databases were followed in the 1980s by DNA sequence databases in the USA (GenBank, [37]), Europe (EMBL, [38]) and Japan (DDBJ, [39]). In the beginning, these databases could be searched only by text queries, but later versions included the possibility to enter a sequence query and search the database using a sequence comparison algorithm. This was done using a pairwise sequence comparison. In pairwise sequence comparisons, only two genes or proteins are compared to each other. A second method of sequence comparison is the multiple sequence alignment, through which complete gene families can be aligned to each other.

1.3.2 Pairwise sequence comparison

Pairwise sequence comparisons have been widely used since 1970, when Needleman and Wunsch invented their Needleman-Wunsch algorithm [40]. This algorithm is a ‘global alignment’ algorithm, which means that the two input sequences are being aligned to each other completely, even when there are parts in the sequences that are very difficult to align. A different and more popular way of sequence alignment is the ‘local alignment’, where only stretches of the two sequences are being aligned. A local alignment algorithm stops aligning when it reaches the point where the two sequences cannot be aligned anymore in a right way. The first local alignment algorithm was the Smith-Waterman algorithm [41], created in 1981. Both the Needleman-Wunsch and the Smith-Waterman algorithm make use of ‘dynamic programming’, a concept from computer science that is very applicable to biological systems. It has the desirable property that it is guaranteed to find the optimal local alignment with respect to the scoring system being used. The main difference of the Smith-Waterman algorithm to the Needleman-Wunsch algorithm is that negative scoring matrix cells are set to zero, which renders the local alignments visible. The motivation for local alignment is the difficulty to obtain correct alignments in regions of low similarity between distantly related biological sequences, because mutations have added too much 'noise' in evolutionary times to allow for a meaningful comparison of these regions. Local alignment avoids these regions altogether and focuses on those with a positive score, i.e. those with an evolutionary conserved signal of similarity.

A large disadvantage of the Smith-Waterman algorithm is that it is fairly compute intensive. As a result, it has largely been replaced in practical use by faster heuristic algorithms such as FASTA [42] and BLAST [43]. Although not guaranteed to find optimal alignments, these are much more efficient, which explains their popularity over the past 15 years. However, since compute power is doubling every two years according to

Moore's law [44], the Smith-Waterman algorithm is gaining popularity again, despite its greedy character. We used the Smith-Waterman implementation to create an all-against-all sequence comparison database named Protein World [45], which was the main dataset for our ortholog benchmarking study mentioned in the previous paragraph.

1.3.3 Multiple sequence alignment

As has been mentioned, some ortholog identification methods require the calculation intensive steps of multiple sequence alignment and phylogenetic tree building to be done. In a multiple sequence alignment more than two sequences can be aligned, making possible identification of conserved sequences (motifs). The most similar sequences are aligned first, after which the less related sequences are added successively to the alignment until the entire query set has been incorporated into the solution.

The most widely used program for doing multiple sequence alignments has, for a long time, been ClustalW [46]. This algorithm consists of three main stages: (I) all pairs of sequences are aligned separately in order to calculate a distance matrix giving the divergence of each pair of sequences; (II) a guide tree is calculated from the distance matrix; (III) the sequences are progressively aligned according to the branching order in the guide tree. Alignment can be achieved by either fast approximate alignments or full dynamic programming for the distance calculations used to make the guide tree. The past years several alternatives for ClustalW have been created, such as T-Coffee [47], MAFFT [48] and MUSCLE [49].

T-Coffee [47] is slower than ClustalW but generally produces more accurate alignments for distantly related sequence sets. T-Coffee uses the output from ClustalW as well as another local alignment program LALIGN, which finds multiple regions of local alignment between two sequences. The resulting alignment and phylogenetic tree are used as a guide to produce new and more accurate weighting factors.

MAFFT (Multiple Alignment by Fast Fourier Transform) [48] is a very fast alternative to ClustalW and T-Coffee. MAFFT includes two novel techniques: (I) homologous regions are rapidly identified by the fast Fourier transform (FFT), in which an amino acid sequence is converted to a sequence composed of volume and polarity values of each amino acid residue; (II) a simplified scoring system that performs well for reducing CPU time and increasing the accuracy of alignments even for sequences having large insertions or extensions as well as distantly related sequences of similar length.

MUSCLE (MUltiple SequenCe alignment by Log-Expectation) [49] is a so-called iteration-based multiple alignment program, because it repeatedly realigns the initial sequences as well as adding new sequences to the growing multiple sequence alignment. Progressive methods such as ClustalW are strongly dependent on a high-quality initial alignment: once a sequence has been aligned into the multiple sequence alignment, its alignment is not considered further. This approximation improves efficiency at the cost of accuracy. By contrast, iterative methods can return to previously calculated pairwise alignments or sub-MSAs incorporating subsets of the query sequence as a means of optimizing a general objective function such as finding a high-quality alignment score.

A multiple alignment is often followed by the creation of a phylogenetic tree, from which evolutionary relationships (such as orthology, paralogy and xenology) between the studied proteins or genes can be inferred. Figure 4 shows a part of a multiple alignment of human toll-like receptors, together with a phylogenetic tree (both created by ClustalW).

```
TLR1 --SIP-KQVVKLEALQELNVAFNSLTDLPG--CGSFSSLSVLIIDHNSVSHPSADFFQ-S

TLR2 --LFS----LNLPQLKELYISRNKLMTLPD--ASLLPMLLVLKISRNAITTFSKEQLD-S

TLR3 --EIPVEVFKDLFELKIIDLGLNNLNTLPASVFNNQVSLKSLNLQKNLITSVEKKVFGPA

TLR4 --QLSPTAFNSLSSLQVLNMSHNNFFSLDTFPYKCLNSLQVLDYSLNHIMTSKKQELQHF

TLR5 ETELCWDVFEGLSHLQVLYLNHNYLNSLPPGVFSHLTALRGLSLNSNRLTVLSHNDLP--

TLR6 --SVP-KQVVKLEALQELNVAFNSLTDLPG--CGSFSSLSVLIIDHNSVSHPSADFFQ-S

TLR7 --SFSWKKLQCLKNLETLDLSHNQLTTVPERLSNCSRSLKNLILKNNQIRSLTKYFLQ-D

TLR8 --FFNWTLLQQFPRLELLDLRGNKLLFLTDSLSDFTSSLRTLLLSHNRISHLPSGFLS-E

TLR9 --FFKWWSLHFLPKLEVLDLAGNQLKALTNGSLPAGTRLRRLDVSCNSISFVAPGFFS-K

TLR10--TVP-KETIHLMALRELNIAFNFLTDLPG--CSHFSRLSVLNIEMNFILSPSLDFVQ-S
        .       :  *. : :  * :  :          *  *  . * :       .
```

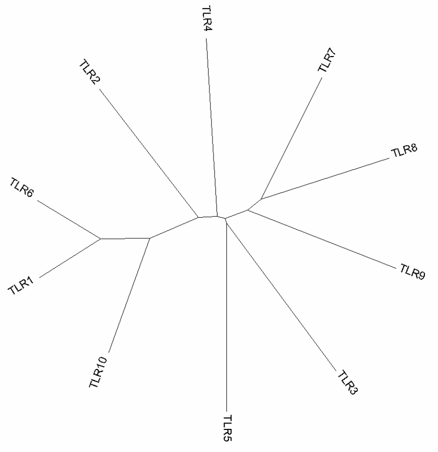


Figure 4. Example of multiple alignment and phylogenetic tree
Part of ClustalW multiple alignment (left) and phylogenetic tree (right) of human toll-like receptors.

1.4 Phylogeny

1.4.1 Phylogeny of genes/proteins

As shown in the previous paragraph, building a phylogenetic tree is a logical step after doing a multiple alignment. Phylogenetics is the field concerning the study of evolutionary relatedness, either among groups of organisms or among groups of genes/proteins. The phylogenetic tree in figure 4 shows the evolutionary relations between groups of proteins. It is generated by ClustalW using the neighbor-joining method [50]. This method is based on the minimum evolution criterion for phylogenetic trees, i.e. the topology that gives the least total branch length is preferred at each step of the algorithm. However, neighbor-joining may not find the true tree topology with least total branch length because it is a greedy algorithm that constructs the tree in a step-wise fashion. Even though it is sub-optimal in this sense, it has been extensively tested and usually finds a tree that is quite close to the optimal tree. Another method for calculating phylogenetic trees from multiple alignment is the UPGMA method [51], but this method is not as good as the neighbor-joining method in terms of minimizing the effects of unequal evolutionary rates in different lineages and giving better estimates of individual branch lengths.

1.4.2 Phylogeny of species

Phylogenies can also be made for species or groups of species. The field of evolutionary relationships between organisms, and their classification, is usually referred to as ‘taxonomy’. One of the most popular taxonomy databases is NCBI taxonomy [34]. In the past the classification was simply done by looking at an organism’s phenotypes. Nowadays the genomics can help in finding the correct ‘tree of life’. For example, species can be compared to each other in the number of shared genes, which is a measure for evolutionary relatedness. Or one can simply look at the level of sequence identity between two organisms: e.g. human and chimpanzee have an overall sequence identity as high as 98%, which makes them very close relatives.

1.4.3 Phylogenetic patterns

One of the advantages of the sequencing of complete genomes, is that one can now see what gene sets occur in which species. This is named ‘phylogenetic occurrence’ and has been incorporated in applications like STRING [12]. The presence or absence of certain genes in certain species can be used to calculate the evolutionary distance between these species: the more similar the ‘phylogenetic pattern’ of species A to the pattern of species B, the smaller the evolutionary distance. These phylogenetic patterns can also be used in a different way: by looking at the patterns of the genes instead of the species (figure 5). This gives much information on the studied gene. For example, if it is present in all species (gene F), the gene is very likely to have an important function. If it is present in only a certain evolutionary branch, it is probably involved in a function that is important in only that branch (genes A, B, E). Genes that have a perfect correlation might be functionally related (genes A, B). Two phylogenetic patterns with a perfect anti-correlation (gene A, B vs. E or gene C vs. D) could be completely different in function, but it could also be possible that they are analogous to each other. Analogous proteins do not have a common ancestor (i.e. are not homologous) but they are (like orthologs) performing the same function in different species.

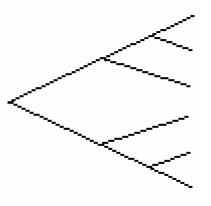
		Gene A	Gene B	Gene C	Gene D	Gene E	Gene F
	Species A	1	1	1	0	0	1
	Species B	1	1	0	1	0	1
	Species C	1	1	1	0	0	1
	Species D	0	0	0	1	1	1
	Species E	0	0	1	0	1	1
	Species F	0	0	0	1	1	1

Figure 5. Small example of the usage of phylogenetic patterns
Phylogenetic patterns of six genes over a species tree consisting of six species. 1 = present, 0 = absent.

1.5 Drug discovery

1.5.1 Introduction

Drug discovery is the process by which drugs are discovered and/or designed. This process involves the identification of molecular targets or systems, chemical entities which can modify these targets or systems, optimization of structure/function characteristics, pharmacology, toxicology, and finally clinical testing of drug candidates. Once a compound has shown its value in these tests, it will enter the process of clinical trials. The complete drug discovery process as performed nowadays can be divided into two major steps: a research step and a development step (figure 6).

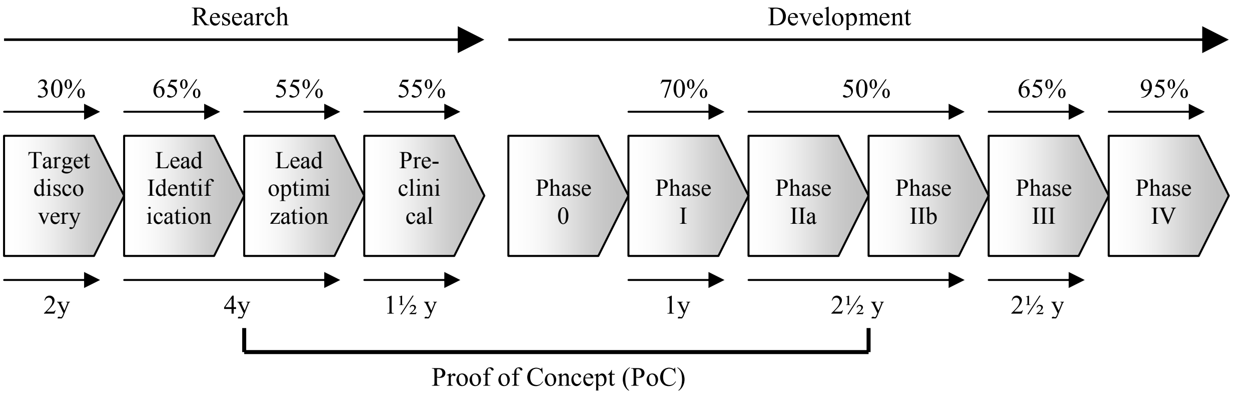


Figure 6. The standard drug discovery pipeline, with the estimated times (in years) [52] and the success rates [53]

The research step itself can be subdivided into four stages: (I) Target discovery: the identification of a biological drug target. This is typically a receptor, enzyme or ion channel that needs to be manipulated to prevent the development of a disease or alleviate symptoms [54]. The availability of new techniques known as genomics and bioinformatics (derived from new knowledge of the human genome) now allows scientists to identify genes coding for potential drug targets of interest. (II) Lead discovery/identification: the development of an assay for the selected target. As new targets are identified, new assays must be developed. Compounds are screened in such assay to find out whether they have any activity on the selected target. This screening is performed by a robotic technology known as 'high throughput screening' (HTS). (III) Lead optimization: compounds found likely to show activity on the target (known as 'hits') are identified as 'lead compounds' and pass to the next stage of lead optimization. Compound properties as potency, selectivity, bioavailability, absorption and metabolism must be assessed. This is achieved by complex functional assays and analytical chemistry. At the same time, lead compounds and their relatives must be synthesized, their molecular structure defined, and their molecular targets constructed. All this is achieved by the use of such advanced techniques including computer assisted drug design and 3D molecular modeling. In addition, the pharmacological profile of the compound is completed and chemical and pharmaceutical feasibility evaluated. (IV) Pre-clinical stage: the drug's pharmacologic and toxic effects are tested in vivo, using animal models.

The development stage consists of several trials on human subjects. Usually, the process is split up into either three or four phases, simply named phase I, phase II and phase III [55], and sometimes an extra phase IV [56]. (I) In phase I a small group (10-80) of healthy volunteers is selected. This phase includes trials designed to assess the safety, tolerability, pharmacokinetics, and pharmacodynamics of a therapy. Phase I trials also normally include dose-ranging studies so that doses for clinical use can be refined. The tested range of doses is usually a small fraction of the dose that causes harm in animal testing (pre-clinical stage). (II) Phase II trials are performed on larger groups (100-300) and are designed to assess clinical efficacy of the therapy. The development process for a new drug often fails during phase II trials due to the discovery of toxic effects or low efficacy. Phase II studies are divided into phase IIA and IIB. Phase IIA is specifically designed to study efficacy, whereas phase IIB is specifically designed to assess dosing requirements. (III) Phase III trials are performed on large patient groups (>1,000) and are aimed at being the definitive assessment of the efficacy of the new therapy, especially in comparison with currently available alternatives. These trials are the most time-consuming, expensive and difficult trials to design and run; especially in therapies for chronic conditions. (IV) Phase IV trials involve the post-launch safety surveillance and ongoing technical support of a drug, mandated by regulatory authorities or undertaken by the sponsoring company for competitive or other reasons. Post-launch safety surveillance is designed to detect any rare or long-term adverse effects over a much larger patient population and timescale than was possible during phase I to III. Such adverse effects detected by Phase IV trials may result in the withdrawal or restriction of a drug.

The above mentioned stages and phases are certainly not rigid; they are almost continuously subject to changes, and different companies use different pipelines. Overmore, some companies use some additional terminology to describe extra phases or umbrella phases. An example of an extra phase is the ‘phase 0’, in which only a small dose of the tested drug is administered to a small number of human subjects (‘microdosing’, [57]). This step takes place just before the other clinical trials (I to IV). An example of an umbrella phase is the ‘Proof of Concept’ or ‘PoC’ phase, which usually refers to the total process of lead optimization, pre-clinical phase, and phase I and IIA trials.

The largest problem in drug discovery is the high attrition rate due to toxicological effects in the phases I to IV, where the costs are extremely high when a trial fails or when the drug is withdrawn from the market, as has happened for example with Vioxx [58] and Pondimin/Redux [59]. A possible solution lies in the application of the new field of translational medicine and therapeutics [3]. This new discipline attempts to more directly connect basic research to the clinical development stages, by early implementation of basic scientific technologies in clinical studies and vice versa. The emphasis is on the linkage between the laboratory and the patient's bedside, often called the ‘bench to bedside’ definition.

1.5.2 Pharmacogenomics

As stated in the first paragraph, genomics can be used to unravel systems in biology, medicine and in the field of drug discovery, which can be useful for fundamental science but can also be applied in industry. The combination of genomics and pharmaceutics is named ‘pharmacogenomics’ or, in short, ‘PGx’ [60] and has an application in medicine and industry [61]. Different definitions of pharmacogenomics exist, but one of the most used definitions states that “pharmacogenomics seeks to apply the field of genomics to improve the efficacy and safety of therapeutics” [62]. In other words: “pharmacogenomics uses genome-wide approaches to elucidate the inherited basis of differences between persons in the response to drugs” [63]. Pharmacogenomics researchers try to achieve this improvement in drug safety and efficacy by the discovery of biomarkers. A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention [64]. Biomarkers can be used to validate novel drug targets and predict drug response, thereby reducing attrition of drugs during the clinical phases of drug development. The field of pharmacogenetics, which is often erroneously interchanged with pharmacogenomics, actually refers only to to genetic-based testing to determine patient therapy, in clinical phase I to IV and later stages. An example of the successful implementation of pharmacogenetics is the FDA approved AmpliChip CYP450 by Roche diagnostics, based on cytochrome P450 genotyping [65,66]. Using pharmacogenetics, drug therapy can be optimised with respect to the patient's genotype, to ensure maximum efficacy with minimal adverse effects [67]. Pharmacogenetics is thus part of pharmacogenomics. The past ten years, pharmaceutical companies have been trying to implement pharmacogenomics into their drug discovery pipeline. It is especially useful in the context of genome-wide expression data and their correlation with a drug's efficacy.

1.5.3 Toxicogenomics

Another –genomics field that is important in drug discovery, especially when dealing with adverse effects, is 'toxicogenomics' or 'TGx', which can actually considered to be a part of pharmacogenomics. Toxicogenomics is defined as "the science of using gene expression profiling from xenobiotic-treated cells or tissues to describe and/or predict various toxic outcomes" [68]. Or defined in a different way: "Toxicogenomics studies toxic effects of substances on organisms in relation to the composition of the genome" [69]. Currently, the premier toxicogenomics tools are the DNA microarray and the DNA chip, which are used for the simultaneous monitoring of expression levels of thousands of genes [70]. Toxicogenomics is especially important in the last step of the research stage of the drug discovery process: the pre-clinical step, in which drugs are tested on model organisms.

1.5.4 Genomics and target discovery

Pharmacogenomics and toxicogenomics are used mainly in the later steps of the research stage of the drug discovery process. However, the main application of genomics in drug discovery lies in the very first step of this process: the finding of interesting drug targets. Genomics is important in this 'target discovery' step because of the large number of potential druggable targets in the human genome (the so-called 'druggable genome' [71]): between 2000 and 3000 [72]. This number is based on the sizes of druggable gene families such as G-protein-coupled receptors, nuclear receptors, ion channels and kinases. It is probably still a conservative estimation, because small-molecule targets and antibody targets have not been included. Current drug therapy is based on less than 500 molecular targets [73], which leaves many new drug targets to be found by genomics techniques. Genomics-based discovery of novel drug targets is however dependent on a good hypothesis that can be addressed through experimental manipulation and the availability of good biological models [74].

1.6 Application of orthology in drug discovery

1.6.1 Model organisms

The concept of model organisms emerged in the beginning of the last century [75], in which mouse, corn, *Drosophila* and *Paramecium* were studied. These species were followed in the forties by species such as *S. cerevisiae*, *E. coli*, *C. elegans* and *A. thaliana*. The impact of model organisms was large, especially in the field of genetics, in which many scientists were thinking about their research organisms as representatives of living things in general. Now we know that this is not a realistic view, since the differences between species won't allow extrapolating research findings from one species to all species. However, since the sequencing of model organisms such as fruit fly [76], worm [77], mouse [78], rat [79], and chicken [80] and the completion of the Human Genome Project [81], we know that the level of similarity on the genome level between man and model organisms can be very high. This similarity is supported by some experimental evidence: many signaling pathways that were once thought to define humans are actually conserved in model organisms. A *C. elegans* nematode placed on the antidepressant fluoxetine has increased serotonin levels in its brain [82]. The set of genes responsible for Alzheimer's disease in mouse seems to be similar to those in man [83]. Many biological processes, including those that are relevant to human diseases, are highly conserved between humans and *Drosophila* [84]. In the past, model organisms were picked to study specific biological systems, like neural development in *C. elegans* and inheritance in *D. melanogaster* [75]. Nowadays, they are utilized in a much broader sense because of the completed sequencing of their genomes. New model organisms are chosen for their evolutionary importance (e.g. *C. intestinalis*, the vertebrate ancestor), which makes them suitable for comparative genomics studies, or because of their importance to society (e.g. rice). Another important use of model organisms lies in the testing of drugs by pharmaceutical companies, in stage IV of the research phase of the drug discovery pipeline.

1.6.2 Pharmacophylogenomics

In spite of the high genetic similarity between man and some model organisms, there is, of course, no species that is the perfect model organism for man. Sometimes a drug that works excellent in a model organism, does not work at all in humans. Or the drug can have several side effects that are absent in the model organism. The concept of orthology plays an important role in understanding why these interspecies differences in response to a certain drug occur. By determining orthologous genes and proteins between the model organism and human, we are able to map the molecular pathways in both species in which this drug is present. This recent combination of phylogenomics and pharmacology has been named 'pharmacophylogenomics' [85]. In the past several pathways, like the citric-acid cycle [86], have been studied successfully in multiple genomes, creating an understanding of the cross-species differences in these pathways. Databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG) [87] enable researchers to perform analyses like these in a straightforward manner. The KEGG database makes use of orthology to map one or more proteins from species A to one or more proteins in species B. This orthology determination, however, can be performed in several ways. The quality of the orthology identification determines if any interspecies differences in response to a drug can be explained successfully.

The major goal of pharmacophylogenomics is to reduce the attrition rate of the drug discovery process. This reduction can be obtained by offering guidelines regarding the druggability of targets: by looking at the evolutionary history of a target, it can suggest a number of possible target 'property filters', e.g. degree and nature of paralogy, breadth of expression, interaction potential and evolutionary rates [85]. This way, pharmacophylogenomics should be able to increase the predictive value of pre-clinical studies. Pharmacophylogenomics is thus applied mainly in the early stages of the research phase (figure 6), when rejecting a compound costs much less time and money than when it is already in the development phase. Besides reducing the attrition rate, it tries to shorten the whole drug discovery pipeline; new genomics techniques should be able to make this timeline shorter than the currently common fourteen years [1]. It must be made clear, however, that the field of pharmacophylogenomics is still in it earliest stage. It certainly holds a promise, but it still needs to prove its value in the upcoming years.

1.7 Thesis outline

This thesis focuses on the concept of orthology, its methodology and its applications.

In **chapter 2** we compare the quality of six popular methods for the identification of orthologous proteins. We use the rule that orthologous proteins should have highly similar functions, and compare these methods by using functional genomics data, such as expression data, protein interactions, protein domains and gene order. This gives a clear view of what ortholog identification method to use in what case: some methods are more sensitive but less selective, while others are more selective but less sensitive.

In **chapter 3** some pairwise sequence comparison methods and statistical significance scores are tested on how well they predict structural similarities. We compare the Smith-Waterman implementations SSEARCH [88], Biofacet [89], ParAlign [90] and Paracel [91], as well as FASTA [42] and BLAST [43]. The Biofacet z-score is compared with the e-value of the other algorithms. This makes clear what sequence comparison method and what statistical significance score to use, taking into account the quality of the method as well as time limiting factors.

In **chapter 4** we present a web application named PhyloPat that gives the user the possibility to input a phylogenetic pattern and check which genes have that phylogenetic pattern. The database behind the application has been constructed using a set of 21 fully sequenced genomes, available through the Ensembl [92] database. The single linkage clustering based on many-to-many orthologous relationships provided by Ensembl is very accurate and reliable, and creates phylogenetic clusters that can be used for more kinds of evolutionary research than phylogenetic pattern studies alone. For example, it makes possible the study of expansions or deletions of certain phylogenetic lineages.

Chapter 5 discusses a study on the evolution of the immune system from model organism to man. This includes all proteins that are in some way connected to the immune system in man and model organisms such as macaque, mouse, rat, dog and chicken.

In **chapter 6** we will have a look at the evolutionary dynamics of bidirectional gene pairs, i.e. two genes sharing a promoter sequence that is lying in between the two genes in the genome. These bidirectional (or head-to-head) gene pairs are remarkably abundant in the human genome: much more than 25% of the human gene pairs have a head-to-head orientation. If there would not be any orientation preference, there would be 25% head-to-head genes, 50% head-to-tail/tail-to-head genes and 25% tail-to-tail genes. Using orthology, we check if these head-to-head gene pairs are still lying in the same orientation in other vertebrate genomes. Furthermore, we check if these head-to-head genes are in close proximity to each other. The sharing of a promoter sequence is much more likely when there is only 600-1000 bp in between the two genes.

Chapter 7 discusses the cross-species connection of transcriptional units, i.e. groups of EST and mRNA sequences that actually belong to one single gene. These gene oriented sequence clusters (transcriptional units) will provide possibilities to identify alternative transcription, SNPs or sequencing errors if sufficient sequences are available. We present a set of algorithms that allows construction of these transcriptional units. To compare the transcriptional units from different species, we need to connect them by using ortholog identification methods such as best bidirectional hit. We compare our set of transcriptional units to Unigene clusters to see which approach is best, and check if our method still needs to be improved.

Chapter 8 gives a short discussion to this thesis, with some concluding remarks.

1.8 References

1. Myers S, Baker A: Drug discovery--an operating model for a new era. *Nat Biotechnol* 2001, 19(8):727-730.
2. Kola I, Landis J: Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 2004, 3(8):711-715.
3. Fitzgerald GA: Opinion: anticipating change in drug development: the emerging era of translational medicine and therapeutics. *Nat Rev Drug Discov* 2005, 4(10):815-818.
4. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 1977, 265(5596):687-695.
5. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC: The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* 2006, 34(Database issue):D332-334.
6. Wikipedia: Genomics [http://en.wikipedia.org/wiki/Genomics]
7. PubMed [http://www.pubmed.gov/]
8. Altman RB, Klein TE: Challenges for biomedical informatics and pharmacogenomics. *Annu Rev Pharmacol Toxicol* 2002, 42:113-133.
9. Kim JH: Bioinformatics and genomic medicine. *Genet Med* 2002, 4(6 Suppl):62S-65S.
10. Kriventseva EV, Fleischmann W, Zdobnov EM, Apweiler R: CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res* 2001, 29(1):33-36.
11. Reichard K, Kaufmann M: EPPS: mining the COG database by an extended phylogenetic patterns search. *Bioinformatics* 2003, 19(6):784-785.
12. Snel B, Lehmann G, Bork P, Huynen MA: STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 2000, 28(18):3442-3444.

13. Richardson MK, Hanken J, Gooneratne ML, Pieau C, Raynaud A, Selwood L, Wright GM: There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anat Embryol (Berl)* 1997, 196(2):91-106.
14. Arthur W: The emerging conceptual framework of evolutionary developmental biology. *Nature* 2002, 415(6873):757-764.
15. Hardison RC: Comparative genomics. *PLoS Biol* 2003, 1(2):E58.
16. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2003, 2(2):13.
17. Cracraft J: Phylogeny and evo-devo: characters, homology, and the historical analysis of the evolution of development. *Zoology (Jena)* 2005, 108(4):345-356.
18. Davison D: Sequence similarity ('homology') searching for molecular biologists. *Bull Math Biol* 1985, 47(4):437-474.
19. Fitch WM: Distinguishing homologous from analogous proteins. *Syst Zool* 1970, 19(2):99-113.
20. Li WH, Yang J, Gu X: Expression divergence between duplicate genes. *Trends Genet* 2005, 21(11):602-607.
21. Mirny LA, Gelfand MS: Using orthologous and paralogous proteins to identify specificity determining residues. *Genome Biol* 2002, 3(3):PREPRINT0002.
22. Gray GS, Fitch WM: Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. *Mol Biol Evol* 1983, 1(1):57-66.
23. Gabaldon T, Huynen MA: Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci* 2004, 61(7-8):930-944.
24. Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. *Science* 1997, 278(5338):631-637.
25. Devos D, Valencia A: Practical limits of function prediction. *Proteins* 2000, 41(1):98-107.
26. Bork P, Koonin EV: Predicting functions from protein sequences--where are the bottlenecks? *Nat Genet* 1998, 18(4):313-318.
27. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ et al: The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* 2005, 33(Database issue):D284-288.
28. Remm M, Storm CE, Sonnhammer EL: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001, 314(5):1041-1052.
29. Alexeyenko A, Tamas I, Liu G, Sonnhammer EL: Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* 2006, 22(14):e9-e15.
30. Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000, 28(1):33-36.
31. Li L, Stoeckert CJ, Jr., Roos DS: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003, 13(9):2178-2189.
32. Dehal PS, Boore JL: A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* 2006, 7:201.

33. Jothi R, Zotenko E, Tasneem A, Przytycka TM: COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics* 2006, 22(7):779-788.
34. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S et al: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2006, 34(Database issue):D173-180.
35. Sanger F, Tuppy H: The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem J* 1951, 49(4):481-490.
36. Dayhoff MO: Atlas of protein sequence and structure, vol. 5. Washington, D.C: National Biomedical Research Foundation, Georgetown University; 1972.
37. Burks C, Fickett JW, Goad WB, Kanehisa M, Lewitter FI, Rindone WP, Swindell CD, Tung CS, Bilofsky HS: The GenBank nucleic acid sequence database. *Comput Appl Biosci* 1985, 1(4):225-233.
38. Hamm GH, Cameron GN: The EMBL data library. *Nucleic Acids Res* 1986, 14(1):5-9.
39. Tateno Y, Gojobori T: DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Res* 1997, 25(1):14-17.
40. Needleman SB, Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970, 48(3):443-453.
41. Smith TF, Waterman MS: Identification of common molecular subsequences. *J Mol Biol* 1981, 147(1):195-197.
42. Pearson WR, Lipman DJ: Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 1988, 85(8):2444-2448.
43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215(3):403-410.
44. Moore G: Cramming more components onto integrated circuits. In: *Electronics Magazine*. 1965.
45. Protein World [http://www.cmbi.ru.nl/pw/]
46. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994, 22(22):4673-4680.
47. Notredame C, Higgins DG, Heringa J: T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 2000, 302(1):205-217.
48. Katoh K, Misawa K, Kuma K, Miyata T: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002, 30(14):3059-3066.
49. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004, 32(5):1792-1797.
50. Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987, 4(4):406-425.
51. Jones JH, Lennard-Jones JE, Morson BC, Chapman M, Sackin MJ, Sneath PH, Spicer CC, Card WI: Numerical taxonomy and discriminant analysis applied to non-specific colitis. *Q J Med* 1973, 42(168):715-732.
52. Folkertsma S: The nuclear receptor ligand-binding domain: from biological fucntion to drug design - a protein family-based approach. Nijmegen; 2006.

53. Nwaka S, Ridley RG: Virtual drug discovery and development for neglected diseases through public-private partnerships. *Nat Rev Drug Discov* 2003, 2(11):919-928.

54. Drug discovery: a challenging process [http://www.organon.com/innovations/process/discovery/]

55. What is a clinical trial? [http://www.nci.nih.gov/clinicaltrials/learning/what-is-a-clinical-trial]

56. Hakkarainen H, Hattab JR, Venulet J: Phase IV research by pharmaceutical companies. *Pharmacopsychiatry* 1984, 17(5):168-175.

57. Garner RC, Lappin G: The phase 0 microdosing concept. *Br J Clin Pharmacol* 2006, 61(4):367-370.

58. Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, Day R, Ferraz MB, Hawkey CJ, Hochberg MC et al: Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *N Engl J Med* 2000, 343(21):1520-1528, 1522 p following 1528.

59. Connolly HM, Crary JL, McGoon MD, Hensrud DD, Edwards BS, Edwards WD, Schaff HV: Valvular heart disease associated with fenfluramine-phentermine. *N Engl J Med* 1997, 337(9):581-588.

60. Salerno RA, Lesko LJ: Three years of promise, proposals, and progress on optimizing the benefit/risk of medicines: a commentary on the 3rd FDA-DIA-PWG-PhRMA-BIO pharmacogenomics workshop. *Pharmacogenomics J* 2006, 6(2):78-81.

61. Marshall A: Genset-Abbott deal heralds pharmacogenomics era. *Nat Biotechnol* 1997, 15(9):829-830.

62. Mehr IJ: Preparing for the revolution--pharmacogenomics and the clinical lab. *Pharmacogenomics* 2000, 1(1):1-4.

63. Evans WE, McLeod HL: Pharmacogenomics--drug disposition, drug targets, and side effects. *N Engl J Med* 2003, 348(6):538-549.

64. Frank R, Hargreaves R: Clinical biomarkers in drug discovery and development. *Nat Rev Drug Discov* 2003, 2(7):566-580.

65. De Leon J: AmpliChip CYP450 test: personalized medicine has arrived in psychiatry. *Expert Rev Mol Diagn.* 2006, 6(3):277-286.

66. Ingelman-Sundberg M: Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. *Trends Pharmacol Sci* 2004, 25(4):193-200.

67. Guo Y, Shafer S, Weller P, Usuka J, Peltz G: Pharmacogenomics and drug development. *Pharmacogenomics* 2005, 6(8):857-864.

68. Barros SA: The importance of applying toxicogenomics to increase the efficiency of drug discovery. *Pharmacogenomics* 2005, 6(6):547-550.

69. Heijne WHM: Toxicogenomics: applications of new functional genomics technologies in toxicology. Wageningen; 2004.

70. Food Standards Agency - Genomics, Transcriptomics and Proteomics: Glossary of Terms [http://www.food.gov.uk/science/ouradvisors/toxicity/cotmeets/49737/49750/49831]

71. Hopkins AL, Groom CR: The druggable genome. *Nat Rev Drug Discov* 2002, 1(9):727-730.

72. Russ AP, Lampel S: The druggable genome: an update. *Drug Discov Today* 2005, 10(23-24):1607-1610.

73. Drews J: Drug discovery: a historical perspective. *Science* 2000, 287(5460):1960-1964.

74. van Duin M, Woolson H, Mallinson D, Black D: Genomics in target and drug discovery. *Biochem Soc Trans* 2003, 31(2):429-432.

75. Davis RH: The age of model organisms. *Nat Rev Genet* 2004, 5(1):69-76.

76. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF et al: The genome sequence of Drosophila melanogaster. *Science* 2000, 287(5461):2185-2195.

77. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 1998, 282(5396):2012-2018.

78. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P et al: Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002, 420(6915):520-562.

79. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE et al: Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 2004, 428(6982):493-521.

80. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME et al: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004, 432(7018):695-716.

81. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al: The sequence of the human genome. *Science* 2001, 291(5507):1304-1351.

82. Schafer WR: How do antidepressants work? Prospects for genetic analysis of drug mechanisms. *Cell* 1999, 98(5):551-554.

83. Morley JE, Farr SA, Kumar VB, Banks WA: Alzheimer's disease through the eye of a mouse. Acceptance lecture for the 2001 Gayle A. Olson and Richard D. Olson prize. *Peptides* 2002, 23(3):589-599.

84. Carroll PM, Fitzgerald K: Model Organisms in Drug Discovery. Hoboken, NJ, USA: John Wiley & Sons, Ltd.; 2004.

85. Searls DB: Pharmacophylogenomics: genes, evolution and drug targets. *Nat Rev Drug Discov* 2003, 2(8):613-623.

86. Huynen MA, Dandekar T, Bork P: Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol* 1999, 7(7):281-291.

87. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 1999, 27(1):29-34.

88. Pearson WR: Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 1991, 11(3):635-650.

89. Codani JJ, Comet JP, Aude JC, Glémet E, Wozniak A, Risler JL, Hénaut A, Slonimski PP: Automatic Analysis of Large-Scale Pairwise Alignments of Protein Sequences. *Methods in Microbiology* 1999, 28:229-244.

90. Rognes T: ParAlign: a parallel sequence alignment algorithm for rapid and sensitive database searches. *Nucleic Acids Res* 2001, 29(7):1647-1652.

91. Paracel [http://www.paracel.com]

92. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T et al: Ensembl 2006. *Nucleic Acids Res* 2006, 34(Database issue):D556-561.

Chapter 2

Benchmarking ortholog identification methods using functional genomics data

Tim Hulsen, Martijn A. Huynen, Jacob de Vlieg and Peter M.A. Groenen

Genome Biology 2006, **7** (4): R31

2.1 Abstract

2.1.1 Background

The transfer of functional annotations from model organism proteins to human proteins is one of the main applications of comparative genomics. Various methods are used to analyze cross-species orthologous relationships according to an operational definition of orthology. Often the definition of orthology is incorrectly interpreted as a prediction of proteins that are functionally equivalent across species, while in fact it only defines the existence of a common ancestor for a gene in different species. However, it has been demonstrated that orthologs often reveal significant functional similarity. Therefore, the quality of the orthology prediction is an important factor in the transfer of functional annotations (and other related information). To identify protein pairs with the highest possible functional similarity, it is important to qualify ortholog identification methods.

2.1.2 Results

To measure the similarity in function of proteins from different species we used functional genomics data, such as expression data and protein interaction data. We tested several of the most popular ortholog identification methods. In general, we observed a sensitivity/selectivity trade-off: the functional similarity scores per orthologous pair of sequences become higher when the number of proteins included in the ortholog groups decreases.

2.1.3 Conclusion

By combining the sensitivity and the selectivity into an overall score, we show that the InParanoid program is the best ortholog identification method in terms of identifying functionally equivalent proteins.

2.2 Background

Orthology is one of the central concepts of comparative genome analysis, but is often misused as a description of functionally equivalent genes in different species. By definition, the term describes the evolutionary relationship between homologous genes whose independent evolution reflects a speciation event, whereas paralogy refers to genes that have diverged from a common ancestor through a gene duplication event [1]. Orthologous genes are more likely to have a functional similarity than paralogous genes, which have often undergone changes in substrate or ligand specificity [2,3]. The high level of functional conservation between orthologous proteins makes orthology highly relevant for protein function prediction. It is also widely used in genome analysis, where the information about a protein in one species is used for the functional annotation of the orthologous protein in another species. At the level of protein-protein interactions, for example, it allows networks of orthologous sequences to be investigated to detect conservation of processes and pathways.

So far, the genomes from more than 200 organisms have been fully sequenced. Of particular interest for medical research are the full genome sequences of human and model organisms, such as fruit fly, worm, mouse, rat, and chicken. Genome sequencing projects on other model organisms, such as the chimpanzee [4], are also close to completion. Identification of orthologous relationships between these model organisms and human allows the functional annotation of a model organism protein to be transferred to its human ortholog.

Given the large amount of data, automated determination of orthology relations is an absolute requirement for an optimal knowledge transfer between the proteins and pathways from different species. Several ortholog identification methods have been described that use sequence comparisons, for example, Clusters of Orthologous Groups (COG) [5], InParanoid [6] and OrthoMCL [7]. One of the most striking differences between the various methods and databases is the level of inclusiveness: the number of proteins from one species that is considered to be part of the same orthologous group. For the best bidirectional hit (BBH) method this number is one, except for theoretical cases where two proteins from species A have the same score to a protein from species B or when one considers fusion or fission of genes [8]. In the euKaryotic Orthologous Groups (KOG) database [9], this number can easily become larger than 100 proteins, for example, for trypsin (KOG3627) in *Homo sapiens*. The reasons for this difference in inclusiveness are twofold. Firstly, there are differences between the algorithms being employed, such as bidirectional best hits, the triangular best-bidirectional hits scheme of the COGs [5], the graph-clustering program OrthoMCL [7], the sequence similarity based InParanoid [6], or a phylogenetic tree algorithm [10]. Secondly, some databases include a wider phylogenetic array of species than others. To give one example, the KOG database [9] aims to include all sequenced eukaryotes. In such a situation, genes resulting from relatively recent gene duplications, like those in the lineage leading to the mammals, will all be part of the same orthologous group. In a database that includes only the mammals, for example, a version of InParanoid that compares mouse and human, these genes will likely be split into different orthologous groups. Comparing only recently diverged species, therefore, allows one to obtain a higher level of evolutionary, and possibly also functional, resolution.

The various published orthology identification methods have led to the recognition that it would be useful to compare these algorithms and use the consistency in the predicted orthologous relations as a measure of reliability [11]. Additionally, several procedures have been proposed to test the reliability of orthology prediction from a single method [6,12]. It has even been proposed that one could actually use functional genomics data to assess the reliability of orthology prediction algorithms to predict functional equivalent genes [13]. However, consistency in the prediction is no measure of statistical or biological significance and the comparison of several ortholog identification methods using functional genomics data is, to the best of our knowledge, a complete new approach to the problem. Here we define and follow a strategy to test the quality of several currently used ortholog identification methods to identify functionally equivalent proteins. Unfortunately, there is no 'gold standard' of protein function that can be used to benchmark ortholog identification methods, as experimentally determined functions are only known for a very small fraction of the proteins in the sequenced genomes. Hence, assessing the quality of different methods currently used is not a straightforward exercise. In our strategy, we use the assumption that functionally equivalent orthologs should behave similarly in functional genomics data [14]. This aspect of conservation of function can be measured in several ways: by similar expression profiles (tissue distribution or regulation), conservation of co-expression, identical domain annotation, conservation of protein-protein interaction or involvement in similar processes (pathways). All of these properties are used here to benchmark the quality of several commonly used ortholog identification methods. The outcome of this benchmark will be useful for determining which ortholog identification method should be used to identify orthologous relationships. Moreover, it gives an idea of which methods are good at predicting different kinds of functional conservation. Some methods appear to be good at predicting conservation of co-expression, while others more accurately predict the conservation of the molecular function. Which ortholog identification method one should use depends on the kind of functional annotation that is to be transferred from one protein to the other. Here we show some examples of the differences between the various kinds of functional conservation in relation to the type of ortholog identification. As a start for building a 'gold standard' of protein function, we also included a comparison with a reference set of 'true orthologs' consisting of five well-studied protein families.

2.3 Results

2.3.1 Direct conservation of functional parameters

First, we measured the conservation of functional parameters between orthologous proteins, examining direct correspondence between human and mouse/worm proteins (figures 1 and 2). This conservation was measured by comparing the expression profiles that provide information about the functional context of a protein (figure 1) and the InterPro accession numbers, which provide information about the molecular function of a protein (figure 2). We determined the correlation in tissue expression patterns between the human-mouse and human-worm orthologous pairs from the six benchmarked methods (figure 1). Note that only proteins for which gene expression data exist are included in this analysis. This is shown by the lower average proteome sizes in, especially, the human-worm analysis, for which it was difficult to map the expression data to the Protein World data. For the human-mouse analysis, this was less difficult. For the three group orthology methods, InParanoid (INP), KOG and OrthoMCL (MCL), a second calculation method was used, which only takes into account the best scoring pair within a group. An examination of only the average correlation shows that the KOG best scoring pair (KOGB) human-mouse set, containing the best scoring human-mouse pair of each KOG, seems to have the highest conservation of function. However, this set has the lowest average proteome size for humanmouse, thus combining a high selectivity with a low sensitivity. If orthology relationships between a larger number of proteins are required, the MCL and MCL best scoring pair (MCLB) sets are good alternatives. Finally, the large standard deviations are a reason to be careful with the interpretation of these results. We do not have this statistical issue when examining the conservation of InterPro accession numbers (figure 2). The ortholog identification methods that create the most orthologous relationships have a larger fraction of equal InterPro accession numbers than the others. The many-to-many non-group methods PhyloGenetic Tree (PGT) and Z 1 Hundred (Z1H) show particularly good scores. Note that these methods use a Smith-Waterman calculation in combination with a Z-value threshold (Monte-Carlo statistics) to define the orthologous relationships (Z ≥ 20 with some additional steps for PGT, Z ≥ 100 for Z1H), whereas the methods with the lower scores, INP, KOG and MCL, use BLAST in combination with E-value statistics.

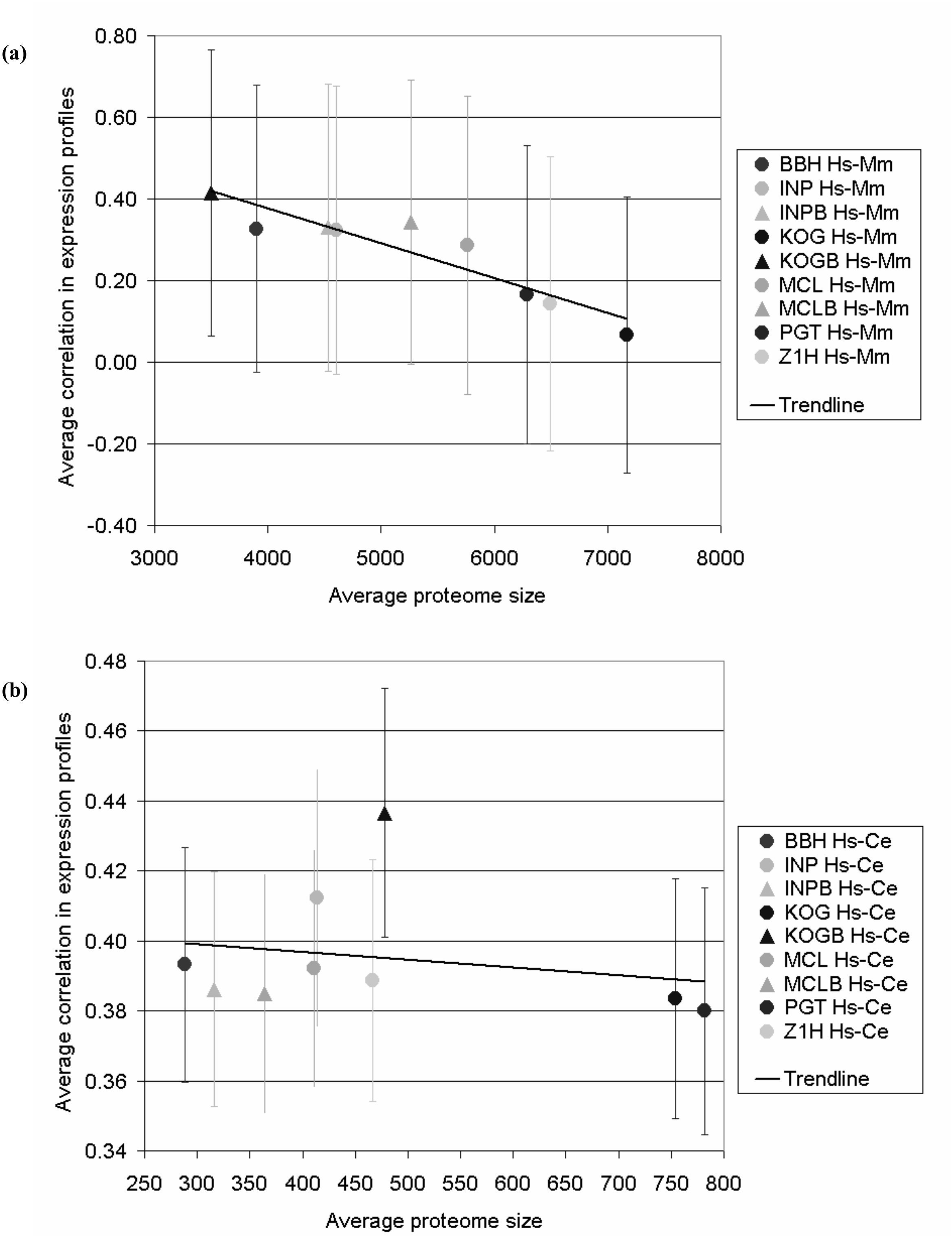


Figure 1. Correlation in expression profiles

Correlation in expression patterns between the **(a)** human-mouse (Hs-Mm) and **(b)** human-worm (Hs-Ce) orthologous pairs from the benchmarked methods versus the average proteome size. Vertical error bars show the standard deviation from the average correlation coefficient. The trendline shown is a linear regression trendline. The methods having a fourth letter 'B' behind the method name, shown as squares in the graph, are group orthology methods in which only the best scoring pairs are taken into account. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*. *Color version on page 147.*

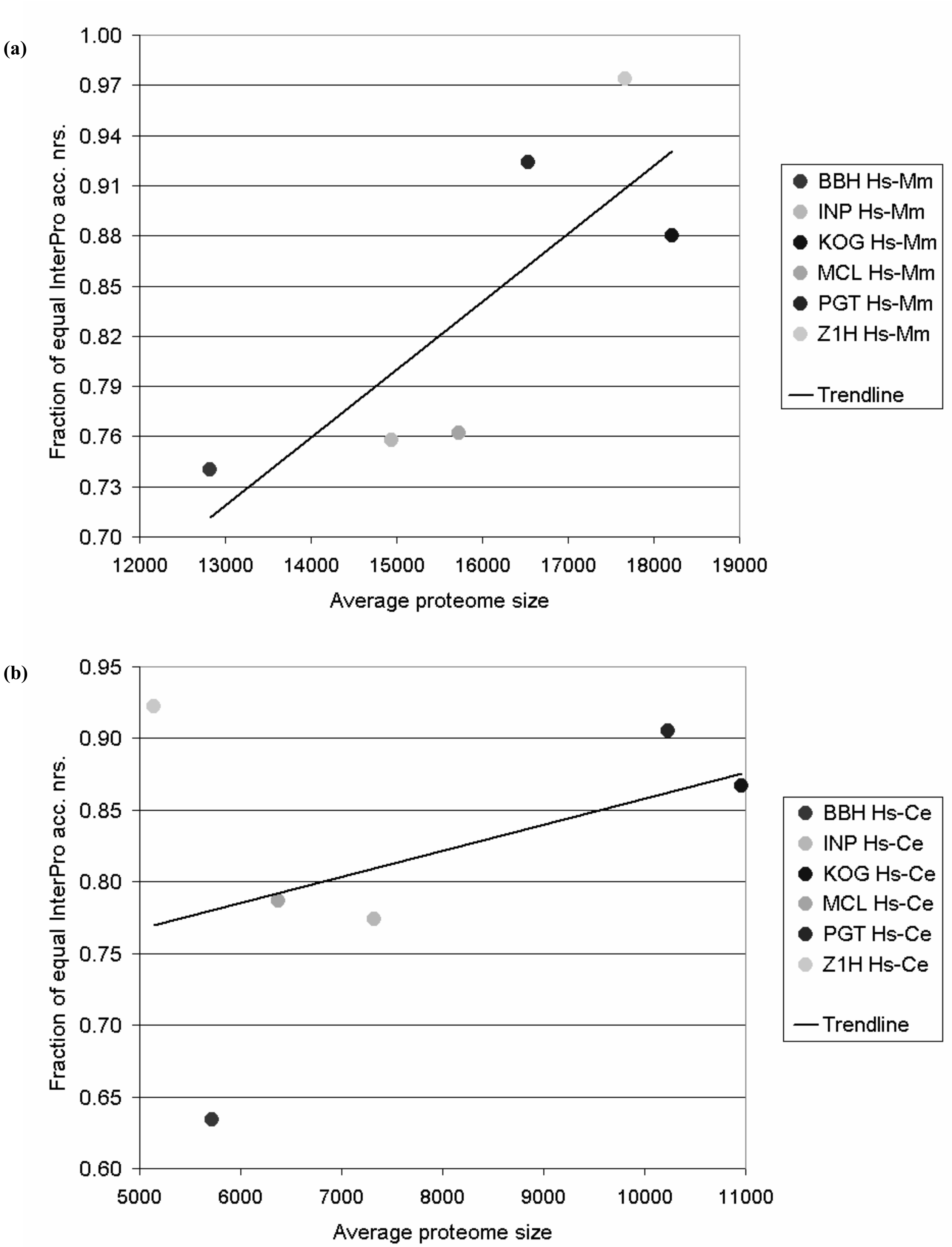


Figure 2. Equal InterPro accession number

Conservation of InterPro accession number between the **(a)** human-mouse (Hs-Mm) and **(b)** human-worm (Hs-Ce) orthologous pairs from the benchmarked methods versus the average proteome size. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*. *Color version on page 148.*

2.3.2 Pairwise conservation of functional parameters

We examined three other methods for orthology prediction benchmarking. In these benchmarks, rather than comparing one-to-one functional correspondence between human and mouse/worm proteins, we compared the correspondence of the relationship between two proteins in human with the relationship between their two orthologs in mouse/worm. In this article, we refer to these methods as 'pairwise conservation of functional parameters' (figures 3, 4 and 5). This functional conservation between two human proteins and two mouse/worm proteins is measured by comparing the co-expression levels (figure 3), the neighboring relationships (figure 4) and the protein-protein interactions (figure 5) between these two species. As described in some recent papers [9,15], the evolutionary conservation of co-expression can be used for function prediction. Here it is used to test which of the ortholog sets can be used to best improve the function prediction, using the Gene Ontology (GO) database [16]. According to our first pairwise benchmark (figure 3), the PGT approach is the best method in the human-mouse analysis, having the highest fraction of equal 4th level GO biological process and the third/fourth largest average proteome. Z1H is the second best method when using conservation of co-expression as a benchmark, having both the second highest sensitivity and the second highest selectivity. The second benchmark, the conservation of gene order, gives completely different results (figure 4): the BBH, INP and MCL methods have the best scores. The three methods with a relatively large average proteome size (PGT, Z1H and KOG) have exceptionally low scores here: all have a fraction of conserved gene order below 0.02. For the conservation of protein-protein interaction (figure 5), the smallest set of all, BBH, has the best score. However, the INP and MCL sets have the best score when both the fraction of conserved protein-protein interaction and the average proteome size are taken into account. Although not as dramatically low as the fractions of conserved gene order, the fractions of conserved protein-protein interaction are still quite low for the three methods with the largest average proteome size.

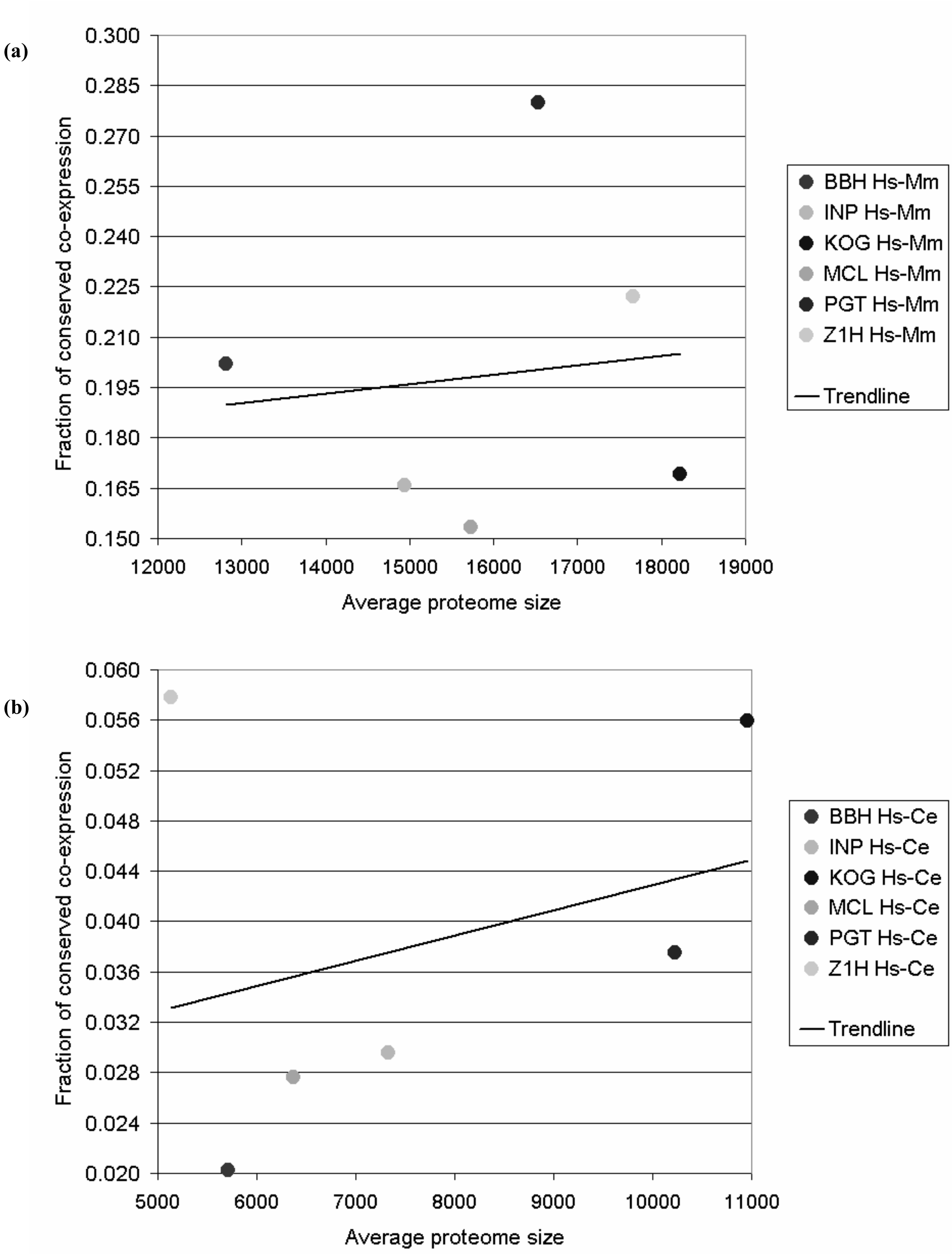


Figure 3. Conservation of co-expression

Conservation of co-expression from human-human gene pairs to orthologous **(a)** mouse-mouse and **(b)** worm-worm gene pairs from the benchmarked methods versus the average proteome size. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*. *Color version on page 149.*

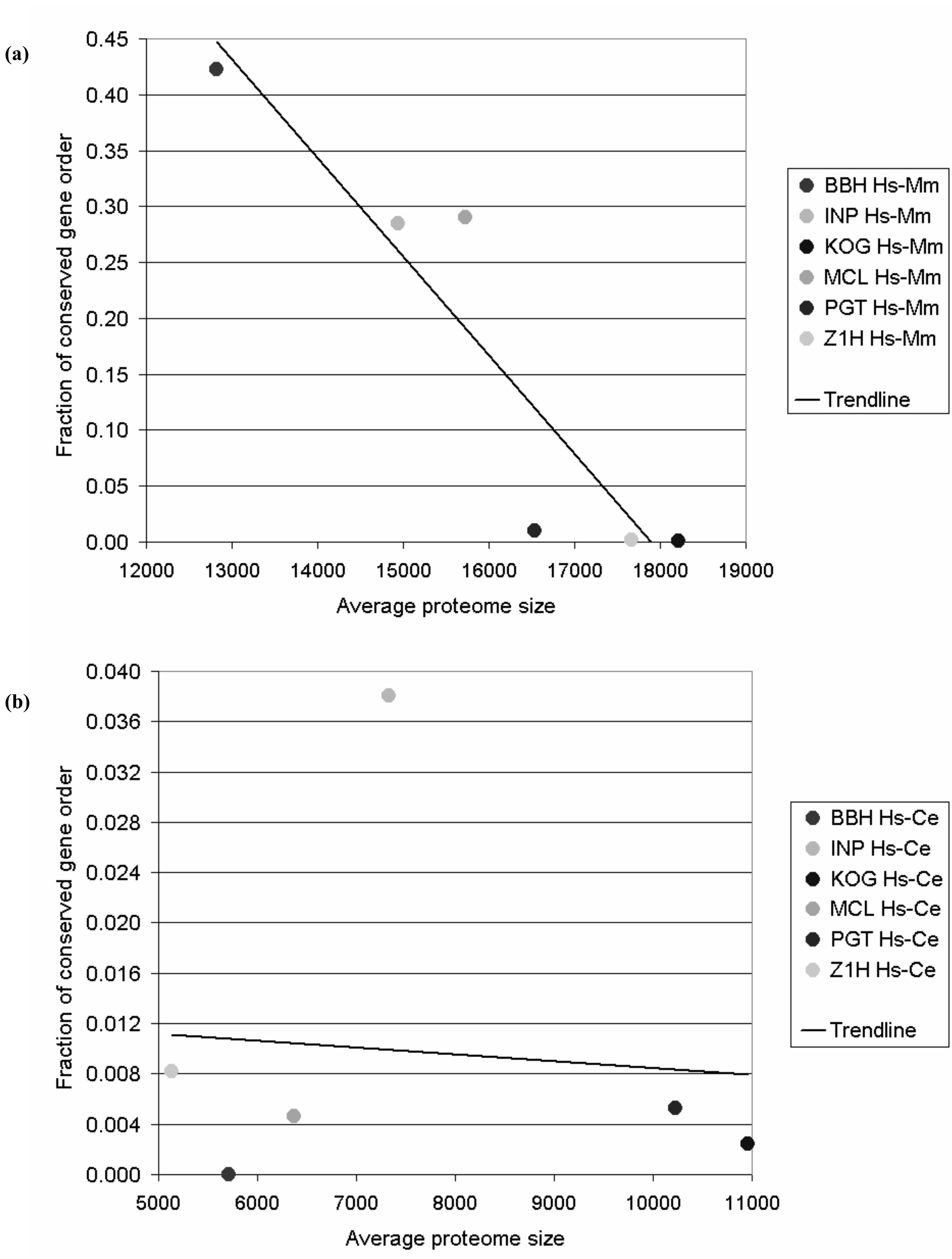


Figure 4. Conservation of gene order

Conservation of gene order from human-human gene pairs to orthologous **(a)** mouse-mouse and **(b)** worm-worm gene pairs from the benchmarked methods versus the average proteome size. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*. *Color version on page 150.*

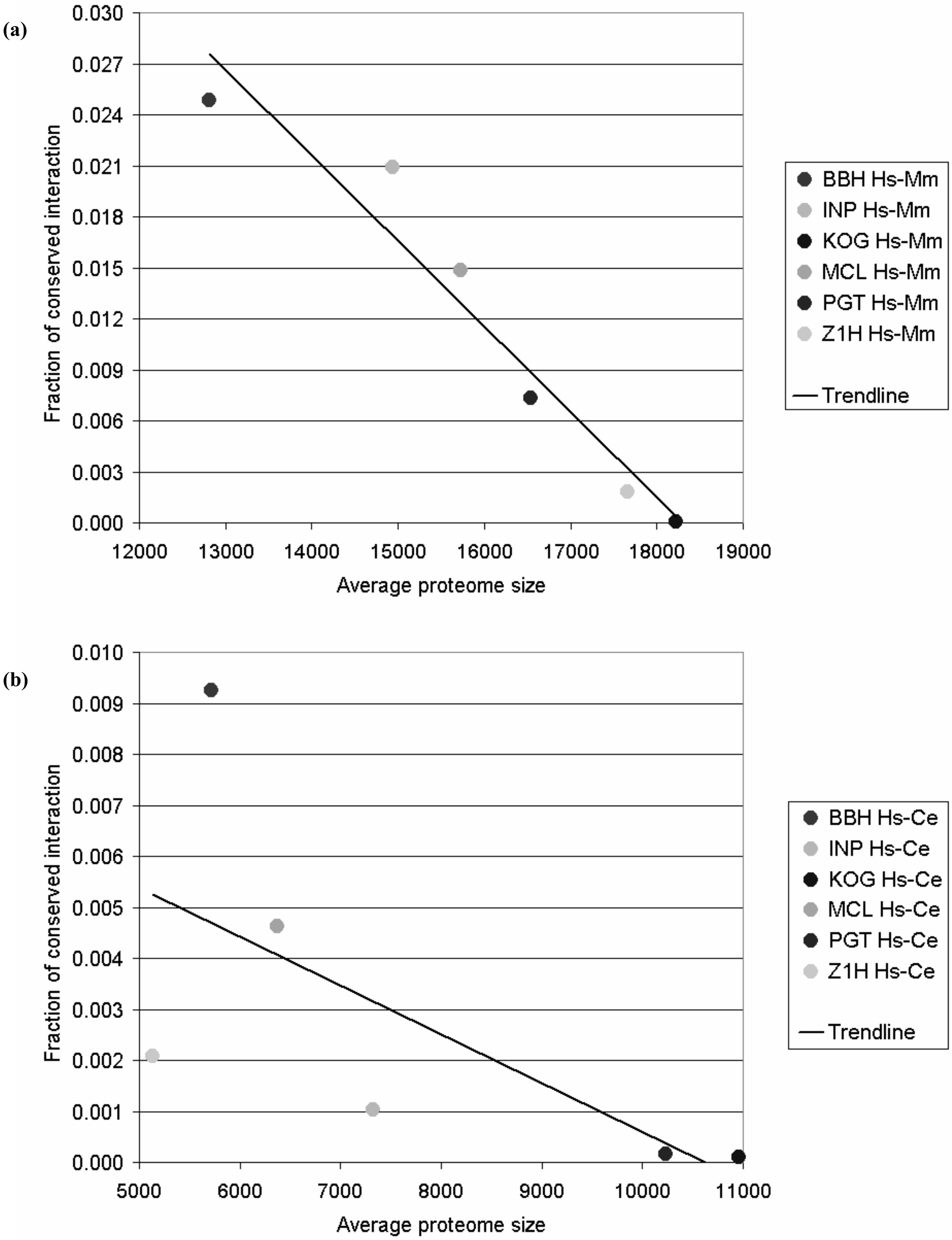


Figure 5. Conservation of protein-protein interaction

Conservation of protein-protein interaction from human-human protein pairs to orthologous **(a)** mouse-mouse and **(b)** worm-worm protein pairs from the benchmarked methods versus the average proteome size. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*. *Color version on page 151.*

2.3.3 Overall results

From the independent results it is difficult to draw a conclusion on which method is best. We therefore determined an overall benchmark of the ortholog identification methods, which are calculated by multiplying the function similarity scores by the average proteome size (table 1). Subsequently, the five resulting scores are combined into one overall score by multiplying them. Each benchmark has its own ranking, on a scale from 1 to 6, and an overall ranking according to the overall score. The overall scores and the overall ranking show that BBH and INP score best, closely followed by MCL. If we combine the several benchmarks into an overall score in a different way, by normalizing all benchmarking scores first (putting the lowest score at 0 and the highest score at 100) and then adding them up, the results are approximately the same (figure 6a for human-mouse). Again, the BBH and INP methods have the best score, followed by the PGT and MCL methods. KOG has a very low overall score. PGT has both a higher score and a larger average proteome size than MCL. The human-worm analysis (figure 6b) shows that the sensitivity/selectivity trade-off is less visible here. The INP method, which has the fourth largest selectivity, has the highest overall score. Z1H, the method with the largest selectivity, has only the second highest score. These results might be influenced, however, by the lower reliability of the human-worm expression data. When combining the results from figure 6a and 6b, we can conclude that the InParanoid algorithm is the best ortholog identification method.

Table 1. Benchmarking scores of ortholog identification methods

Method	Direct conservation of function		Pairwise conservation of function			Overall score
	Co-expression	Equal InterPro accession number	Conservation of co-expression	Conservation of gene order	Conservation of protein-protein interaction	
Hs-Mm						
BBH	1.28E+03 (3)	9.49E+03 (6)	2.59E+03 (4)	5.42E+03 (1)	3.18E+02 (1)	5.42E+16 (2)
INP	1.49E+03 (2)	1.13E+04 (5)	2.48E+03 (5)	4.26E+03 (3)	3.13E+02 (2)	5.57E+16 (1)
KOG	4.73E+02 (6)	1.60E+04 (2)	3.08E+03 (3)	1.42E+01 (6)	1.09E+00 (6)	3.61E+11 (6)
MCL	1.66E+03 (1)	1.20E+04 (4)	2.41E+03 (6)	4.56E+03 (2)	2.34E+02 (3)	5.10E+16 (3)
PGT	1.05E+03 (4)	1.53E+04 (3)	4.63E+03 (1)	1.73E+02 (4)	1.21E+02 (4)	1.56E+15 (4)
Z1H	9.29E+02 (5)	1.72E+04 (1)	3.93E+03 (2)	3.75E+01 (5)	3.17E+01 (5)	7.46E+13 (5)
Hs-Ce						
BBH	2.25E+03 (5)	3.62E+03 (6)	1.16E+02 (6)	0.00E+00 (6)	5.29E+01 (1)	5.00E+10 (6)
INP	3.02E+03 (3)	5.67E+03 (3)	2.17E+02 (4)	2.79E+02 (1)	7.62E+00 (4)	7.90E+12 (1)
KOG	4.20E+03 (1)	9.51E+03 (1)	6.14E+02 (1)	2.64E+01 (5)	1.17E+00 (6)	7.58E+11 (5)
MCL	2.50E+03 (4)	5.01E+03 (4)	1.76E+02 (5)	2.95E+01 (4)	2.94E+01 (2)	1.91E+12 (2)
PGT	3.89E+03 (2)	9.26E+03 (2)	3.84E+02 (2)	5.36E+01 (2)	1.65E+00 (5)	1.22E+12 (4)
Z1H	2.00E+03 (6)	4.74E+03 (5)	2.97E+02 (3)	4.20E+01 (3)	1.07E+01 (3)	1.27E+12 (3)

Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*.

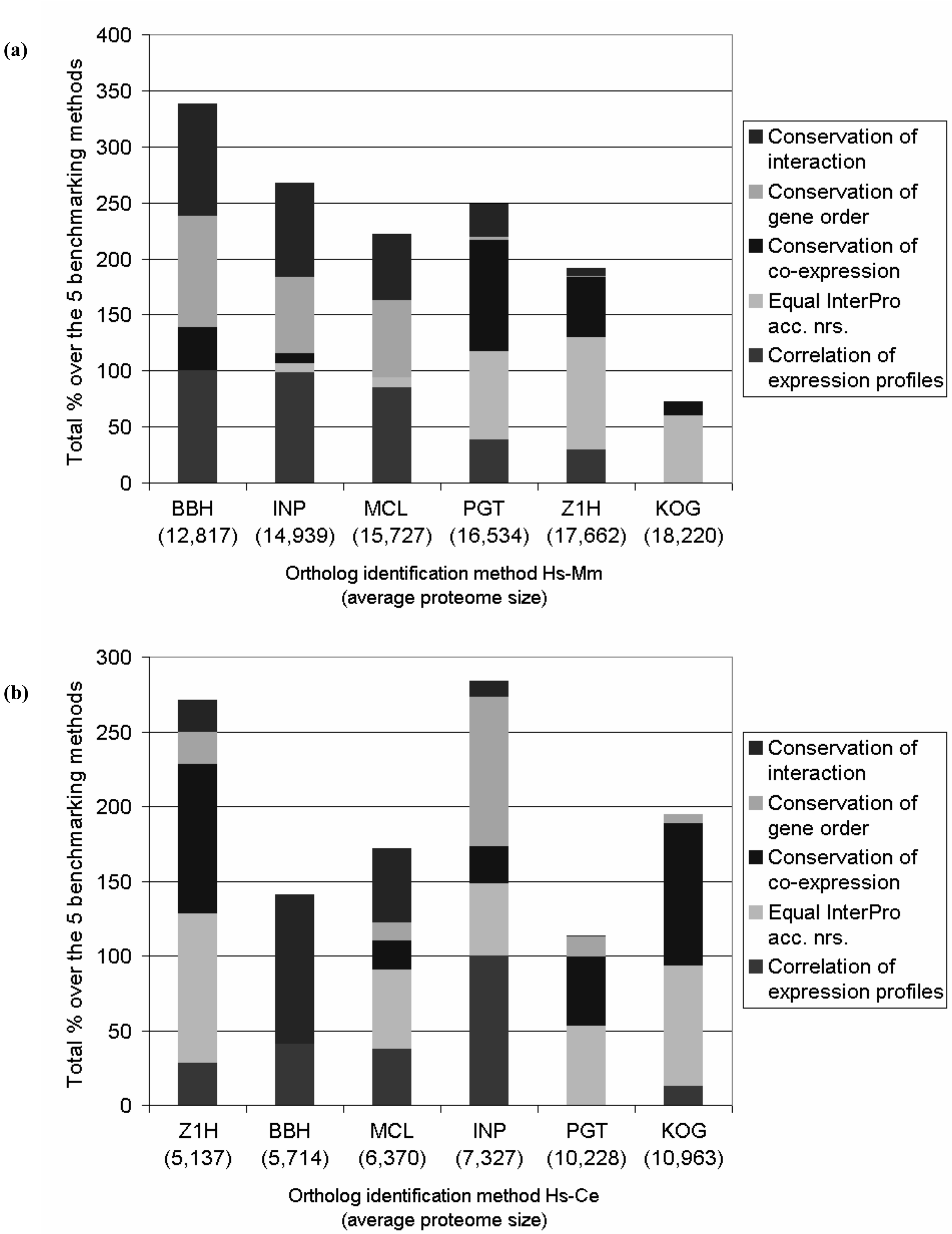


Figure 6. Overall scoring graph

Overall scoring graph, created by adding up all normalized benchmarking scores per ortholog identification method. X-axis, the several ortholog identification methods, sorted by average proteome size or number of protein pairs; Y-axis, the sum of all five benchmarking scores per ortholog identification method. Red, correlation of expression profiles; green, equal InterPro accession numbers; blue, conservation of co-expression; orange, conservation of gene order; purple, conservation of protein-protein interaction. **(a)** Human-mouse (Hs-Mm). **(b)** Human-worm (Hs-Ce). *Color version on page 152.*

2.3.4 Ortholog reference set

We included in our study a 'true ortholog' reference set, consisting of five well-studied protein families: the Hox cluster proteins and hemoglobins (human-mouse), the nuclear receptors and toll-like receptors (human-worm), and the Sm and Sm-like proteins (human-mouse plus human-worm). Table 2 shows the overlap between the orthologs defined by the six different methods and this reference set.

Table 2. Overlap with ortholog reference set

	Method	**Orthologous pairs**	**Orthologous pairs divided by average proteome size**	**False positives**
Hox cluster proteins (Hs, 31 unique proteins; Mm, 35 unique proteins; Hs-Mm, 41 protein pairs)	BBH	26	2.03E-03	3
	INP	28	1.87E-03	3
	KOG	30	1.65E-03	456
	MCL	26	1.65E-03	25
	PGT	33	2.00E-03	350
	Z1H	26	1.47E-03	19
Nuclear receptors (Hs, 22 unique proteins; Ce, 18 unique proteins; Hs-Ce, 29 protein pairs)	BBH	8	1.40E-03	2
	INP	13	1.77E-03	179
	KOG	20	1.82E-03	2,062
	MCL	13	2.04E-03	4
	PGT	11	1.08E-03	180
	Z1H	8	1.56E-03	8
Hemoglobins (Hs, 4 unique proteins; Mm, 9 unique proteins; Hs-Mm, 9 protein pairs)	BBH	2	1.56E-04	2
	INP	6	4.02E-04	8
	KOG	4	2.20E-04	52
	MCL	4	2.54E-04	3
	PGT	4	2.42E-04	23
	Z1H	8	4.53E-04	37
Toll-like receptors (Hs, 10 unique proteins; Ce, 1 unique protein; Hs-Ce, 10 protein pairs)	BBH	0	0	0
	INP	0	0	0
	KOG	10	9.12E-04	1
	MCL	0	0	0
	PGT	5	4.89E-04	86
	Z1H	0	0	0
Sm proteins (Hs, 13 unique proteins; Mm, 17 unique proteins; Hs-Mm, 17 protein pairs)	BBH	5	3.90E-04	8
	INP	5	3.35E-04	8
	KOG	6	3.29E-04	15
	MCL	4	2.54E-04	10
	PGT	7	4.23E-04	18
	Z1H	5	2.83E-04	4
Sm proteins (Hs, 6 unique proteins; Ce, 6 unique proteins; Hs-Ce, 6 protein pairs)	BBH	6	1.05E-03	0
	INP	6	8.19E-04	0
	KOG	4	3.65E-04	1
	MCL	6	9.42E-04	2
	PGT	3	2.93E-04	9
	Z1H	0	0	0

Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*.

The human-mouse Hox cluster proteins are covered best by the PGT method: 33 out of 41 orthologous pairs are detected. The KOG method is the second best with 30 orthologous pairs, and InParanoid is third best with 28 pairs. The other three methods all find the same 26 pairs. However, the KOG and PGT methods also have a high number of false positives. When the number of orthologous pairs is divided by the average proteome size, the BBH method has the highest score, followed by PGT and INP. The nine human-mouse hemoglobin orthologous pairs are almost all detected by the Z1H method. The orthologous pairs/average proteome size ratios of the six different methods do not differ much for this family, which means that the number of detected pairs is proportional to the inclusiveness of the ortholog identification method. PGT and BBH have the best scores when looking at Sm and Sm-like proteins.

As for the human-worm nuclear receptors, the KOG method has the highest number of orthologous pairs. However, KOG has an extremely high number of false positives. When the numbers of orthologous pairs are divided by the average proteome size, the MCL method has the best performance. The Toll-like receptor family, which has only one member in *Caenorhabditis elegans* shows good results for KOG as well, together with the PGT method. For the Sm and Sm-like protein family, the MCL and INP methods have the highest orthologous pairs/average proteome size ratios.

2.4 Discussion

We have tested the quality of a number of ortholog identification methods for protein function prediction by comparing functional genomics data from each of the proteins in a pair identified as orthologs. Orthologs should, in general, have a higher level of function conservation than paralogs. The results show that, in general, the less inclusive the method, the better it performs in terms of function similarity; in other words, there is a certain trade-off between sensitivity and selectivity. We correct for this by taking the function similarity score and multiplying it by the geometric average of the number of unique human proteins and the number of unique mouse/worm proteins within the ortholog set that is being studied (the 'average proteome size'). After multiplying these scores to obtain an overall score (giving each benchmark the same weight), we generate an overall ranking that gives equal weight to both the five different benchmarks and the sensitivity and selectivity. From the results, we conclude that the InParanoid method is the best ortholog identification method. However, some caution should be taken with the overall ranking system. First, the average proteome size now has the same weight as the function similarity score, while one of them might be considered more important than the other. We examined the effect of different weights for these two parameters (1:2 and 2:1 proportions) but did not find any large differences in the results. Second, some benchmarks may produce better results than others, which might be a reason to give different weights to the several benchmarks when combining them into an overall score. For example, the benchmark that uses GO annotations could be less reliable because some of these annotations are actually based on sequence similarity themselves. Third, recent research [17] suggests that the expression levels of physically interacting proteins coevolve. This indicates a strong connection between the third and the fifth benchmark in this study, which could be a reason to leave out one of them. However, coexpression can be the result of processes other than physical interaction only. The differences in the results we got from the two benchmarks also contributed to our decision not to exclude either one of them. Finally, it should be noted that the data we used in our human-mouse analysis was, in general, of higher quality than the data we used in our human-worm analysis. This applies especially to the gene expression data: for the human-mouse set we could use the SNOMED tissue classification, whereas for the human-worm set we found it quite hard to map the tissue samples to each other. The small numbers that were generated in the human-worm analysis also makes this analysis statistically less reliable than the human-mouse analysis.

The conclusion that can be drawn from this study is that the method that should be used to identify orthologs is in fact dependent on the research question one wants to answer using the orthologous relationships. For example, if the goal is to have one or more orthologs for a large number of proteins, one of the methods that allow many-to-many relationships (like InParanoid) should be applied. If selectivity (having as few as possible false positives) is more important than sensitivity (having as many as possible true positives) and having only one ortholog per protein is sufficient, the best bidirectional hit approach should give the best results. Although methods that include phylogenetic inferences to determine phylogenies should, in principle, be the best at establishing orthologous relationships, in practice they suffer from a number of drawbacks that methods solely based on pairwise identities do not have. It is commonplace, for example, to require positions in a sequence alignment to be present in all or most of the sequences in order to use them for deriving a phylogeny with ClustalW. Such requirements drastically reduce the amount of information that can be used to determine orthology relationships. In the absence of easily implementable solutions to this, computational shortcuts like InParanoid give, in our analysis, better results.

Finally, results could differ when different statistical significance scores (unpublished data), scoring matrices, gap penalties, and so on are used for the various alignment algorithms. We tried to minimize the effect of these parameters as much as possible by using the defaults of the several programs, but some programs might still be more suitable for identifying close orthologous relationships than others, while these others might be more appropriate for the identification of distant relationships. The differences observed between our human-mouse (closely related species) and human-worm (distantly related species) analyses support this statement. As for the human-worm analysis, the conservation of functional characteristics and gene order is significantly lower than in human-mouse. The latter is not surprising because millions of years of chromosomal rearrangements during evolution have changed the chromosomal organization significantly. As for the functional aspects, we can conclude that they have been poorly conserved whereas the protein domain organization has been well conserved.

2.5 Conclusion

Because of the high degree of functional similarity between orthologous proteins, the quality of orthology prediction is an important factor in the transfer of functional annotation. To measure the functional similarity of proteins from different species we use functional genomics data, such as protein interaction data and expression data. In general, we observe a sensitivity/selectivity trade-off: the functional similarity scores per orthologous pair become higher when the number of proteins included in the ortholog groups decreases. This trend is more visible in the human-mouse comparison than it is in the human-worm comparison. Presumably, it gets less visible when the phylogenetic distance gets larger. By combining the sensitivity and the selectivity into an overall score, we show that the InParanoid program is the best ortholog identification method in terms of identifying functionally equivalent proteins. The method that should be used to answer a specific research question is, however, also dependent on, for example, the evolutionary distance between the studied species and the desirability of many-to-many orthologous relationships.

2.6 Materials and methods

2.6.1 'Protein World' data set

For an unbiased comparison of all of the covered methods, the same data set was used at all times. This 'Protein World' (unpublished data) data set [18] was created by comparing all of the currently known and predicted proteins (SpTrEMBL [19], RefSeq [20], Ensembl [21]) through the Smith-Waterman algorithm [22], using Z-values to obtain a database-size independent estimate of significance [23]. The Smith-Waterman algorithm has been shown to be more sensitive [24] than its faster (non-dynamic programming) approximations, the BLAST [25] and FASTA [26] algorithms. The data set is freely available through the Center for Molecular and Biomolecular Informatics website [27]. As good expression data and other functional data were available for human, mouse and worm, we used the orthologous relationships between these three species for our study.

2.6.2 Ortholog identification methods

The six ortholog identification methods covered in this study are listed below. Included are the best bidirectional hit method and five many-to-many methods. The many-to-many methods are divided into group orthology methods and non-group orthology methods. The group orthology methods, KOG [9], INP [6] and MCL [7], define several, distinct groups of orthologous genes and proteins. The two many-to-many non-group methods, PGT [10] and Z1H, do not define orthologous groups, but can still determine many-to-many orthologous relationships. Table 3 shows the numbers of orthologous groups, unique proteins and protein pairs within the several ortholog sets. The average proteome size is the geometric average of the total number of unique human proteins and the total number of unique mouse/worm proteins within the determined orthologous relationships.

Table 3. General statistics of ortholog identification methods

Ortholog identification method	Orthologous groups	Protein pairs	Human proteins	Mouse/worm proteins	Average proteome size
Hs-Mm					
BBH	-	12,817	12,817	12,817	12,817
INP	12,610	19,482	15,344	14,545	14,939
KOG	7,874	810,697	20,478	15,640	18,220
MCL	7,002	12,625	16,676*	14,833*	15,727*
PGT	-	85,848	17,302	15,729	16,534
Z1H	-	290,176	19,055	16,149	17,662
Hs-Ce					
BBH	-	5,714	5,714	5,714	5,714
INP	4,135	17,011	9,282	5,784	7,327
KOG	4,155	155,387	12,249	9,812	10,963
MCL	4,705	9,749	7,028	5,774	6,370
PGT	-	49,979	12,499	8,370	10,228
Z1H	-	21,509	6,338	4,163	5,137

*Corrected for Ensembl-SpTrEMBL mapping. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*.

2.6.2.1 Best bidirectional hit

The 'best bidirectional hit' (BBH) method is the most frequently applied method to determine orthologous pairs. It assumes that a cross-species protein pair in which each protein gives back the other protein as being the best hit in the whole other proteome is an orthologous pair. In this research, the best bidirectional hits were determined based on Z-values of the Protein World human-mouse and human-worm set, without a sequence similarity cutoff. In total, 12,817 human-mouse and 5,714 human-worm orthologous pairs were identified. Although the BBH method theoretically can give some many-to-many orthologs, it practically gives only one-to-one orthologous pairs.

2.6.2.2 InParanoid

In the INP method [6], all possible pairwise similarity scores between datasets A-A, B-B, A-B and B-A that score higher than a cutoff (bitscore ≥50, overlap ≥50%) are detected. Then the best bidirectional hits are determined and marked as potential orthologs. The in-species pairs that score higher than these orthologous pairs are marked as additional orthologs. These 'in-paralogs' get confidence values that indicate how similar they are to the main ortholog: 100% is assigned to the main ortholog and 0% is assigned to a sequence with the minimum similarity score required to be marked as in-paralog of a given group. Finally, overlapping groups of orthologs are resolved and bootstrap-based confidence values are added for all groups of orthologs. Additionally, an outgroup proteome can be used to test the significance of the in-paralog scores. InParanoid version 1.35 was downloaded [28] and the program was run using the standard parameters, except for the use of the BLOSUM80 matrix instead of the standard BLOSUM62 matrix. The BLOSUM80 matrix is more appropriate when studying protein pairs with relatively small evolutionary distances. The optional third outgroup proteome was left out. We used Paracel BLAST 1.4.9. Through the INP algorithm, 19,482 orthologous pairs were identified between human and mouse, comprising 12,610 orthologous groups; 17,011 orthologous pairs were identified between human and worm, comprising 4,135 orthologous groups.

2.6.2.3 euKaryotic Orthologous Groups

The KOG database [9] is the eukaryote specific version of the COG database [5]. The latter database is considered by many to be the standard orthology database of this moment. Both the COG and the KOG procedure start with an all-against-all comparison using BLAST, followed by the detection of triangles of mutually consistent, genome-specific best hits (BeTs). Subsequently triangles with a common side are merged to form crude, preliminary KOGs, after which a case-by-case analysis of each candidate KOG is carried out, among others to split fused proteins. The difference between COG and KOG lies within the last step, the manual curation. The KOG procedure pays extra attention to multi-domain proteins, which are quite common in eukaryotes. The KOG database currently consists of seven eukaryotic proteomes. A BLAST all-against-all was used to determine the corresponding KOG for each human, mouse and worm protein within the SpTrEMBL set. Orthologous relationships were determined between all human, mouse and worm proteins within a KOG. Because of the large groups that can be formed by KOGs, no less than 810,697 human-mouse orthologous protein pairs were determined, divided over 7,874 orthologous groups; 155,387 orthologous pairs were identified between human and worm, comprising 4,155 orthologous groups.

2.6.2.4 OrthoMCL

The MCL algorithm [7] starts with an all-against-all BLASTP, after which the reciprocal best similarity pairs between species are marked as putative orthologs and the reciprocal better similarity pairs as recent paralogs. A similarity matrix is calculated, followed by a Markov clustering [29], which determines the orthologous groups. A list of all human and mouse Ensembl protein identifiers linked to an OrthoMCL group ID was obtained from the authors. These Ensembl protein IDs were mapped to the SpTrEMBL proteome using EnsMart [30] version 19.3 [31]. Orthologous relationships were determined between all human and mouse proteins within all 7,002 groups, which gives a total of 12,625 orthologous protein pairs. The loss of defined orthologs was corrected for by calculating how many ensembl IDs mapped to an SpTrEMBL ID (57.3397%). The average proteome size of 9,018 (for human-mouse) was divided by 0.573397, giving a corrected number of proteins of 15,727. The human-worm IDs were obtained through the new OrthoMCL-DB [32]; 9,749 human-worm orthologous protein pairs were identified, comprising procedure 4,705 orthologous groups. Because of the different mapping method, we did not need to correct the human-worm average proteome size.

2.6.2.5 Z 1 Hundred

Within the Z1H method, all cross-species protein pairs that have a Z-score of 100 or higher are considered to be orthologs. The Z-value estimates the statistical significance of a Smith-Waterman dynamic alignment score (SW-score) through the use of a Monte-Carlo process [23]. In this approach, selected pairs of sequences are shuffled randomly 200 times and realigned. The significance of the SW-score of a selected pair is then determined by comparing the SW-score of the selected pair with the scores for the shuffled pairs. By comparing the score with that of the shuffled sequences the method implicitly takes into account effects of sequence composition and sequence length. The Z1H set contains pairs of sequences whose SW-score is a hundred standard deviations higher than the average SW-score for the shuffled sequences. Using the Z1H method, 290,176 human-mouse and 21,509 human-worm orthologous protein pairs were identified. The algorithm does not identify distinct groups of proteins, and is, therefore, a non-group method.

2.6.2.6 PhyloGenetic Tree

The PGT method uses the output generated by multiple alignments and subsequent tree calculation [10] to define orthologous relationships. Although calculations like these are rather time consuming, they should give a better insight into the evolution of the studied proteins and in principle come closest to the original evolutionary definition of orthology. Orthologies were determined by grouping all proteins over the 9 eukaryotic species covered in Protein World that have a Z-value above 20 compared to one of the human proteins, and have a region of homology larger than 50% of the query length. The resulting 23,829 groups were aligned using ClustalW version 1.82 [33], and phylogenies were created using neighbor-joining [34]. For the calculation of the phylogenetic trees we only used the positions that were present in all aligned sequences, and levels of protein sequence identity were translated to evolutionary distances using the Kimura correction as implemented in

ClustalW. The other parameters were set to default. After the calculations, an ortholog identification algorithm selects partitions in the tree that only include orthologs and in-paralogs to define the orthologous relationships per species pair [10]. For human and mouse, 85,848 relationships were identified. For human and worm, 49,979 relationships were identified. Because a phylogenetic tree is calculated for the homologs of every sequence, and the trees are not merged, this method is like the Z1H method, not a pure group method.

2.6.3 Benchmarks

Below are a description and the workflow of the used benchmarks. The first two benchmarks measure 'direct conservation of functional parameters', that is, they examine only one protein in human and one protein in mouse/worm. The last three methods compare the relationship between two proteins in human with the relationship of their two orthologs in mouse/worm ('pairwise conservation of functional parameters').

The results of the group orthology methods were analyzed in two ways: we determined the average score for all pairwise orthology relationships within an orthologous group; and we only considered the best scoring pair within an orthologous group. The latter option obviously leads to a much higher score for the many-to-many orthology relationships. However, by including only one pair of orthologous sequences per orthologous group, that high score is balanced by a reduction in the total number of orthologous relationships (one per orthologous group). Both the number of orthologous relationships and the quality of these relationships are taken into account in the final assessment of the ortholog identification algorithms.

2.6.3.1 Direct conservation of functional parameters

To test the conservation of function, the Pearson correlation between the expression profiles of the proteins in an orthologous pair was calculated. The expression dataset used here [35] was a subset of pathologically normal human and mouse tissue samples from the Gene Logic BioExpress Database product [36]. Because of the small overlap of tissue categories (115 in human, 25 in mouse), the SNOMED [37] tissue categories were used to calculate the correlation coefficient (15 in human, 12 in mouse, 12 overlapping categories). The human dataset consists of 3,269 tissue samples and 44,792 cDNA fragments, the mouse dataset of 859 tissue samples and 36,701 cDNA fragments. A perfect correlation has a score of 1, a perfect anti-correlation has a score of -1. We used expression data from Stuart and colleagues [38] for the human-worm analysis, comparing tissues from both species that had similar expression profiles. For computing time-saving reasons, we used a sample of the dataset to calculate which tissues were similar: the first 10 human tissues were compared with all of the 978 worm tissues, using the first 10 metagenes defined by Stuart et al. The 'best hit' of the worm tissue samples for each human tissue sample was seen as corresponding tissue. These ten corresponding tissues were then used to calculate the Pearson correlation coefficients between the human and worm proteins, from which only the positive correlations were used. Proteome sizes were corrected for this by multiplying them by two, before calculating the average proteome size. For visualization reasons we displayed error bars of only one-eighth of the SD. Because of the differences between the human-mouse and human-worm expression data analyses, we emphasize that the two figures (figures 1a and 1b) should not be compared to each other. The figures can, however, be used to compare the several ortholog identification methods within these species pairs.

The conservation of molecular function can also be benchmarked by examining whether the orthologs are in the same InterPro [39] family. Each InterPro accession number represents a protein family or domain, containing a cross-species set of homologous proteins with its own functional annotation. Proteins within an InterPro protein family have similar domain compositions. Again, the higher the percentage with equal InterPro accession numbers, the better the conservation of function. As InterPro annotation is based on similarity to predefined domains, it is not independent of sequence and cannot be used as a completely independent benchmark. It does, however, allow one to judge to what extent proteins that are regarded as orthologous actually do have the same domain composition. This is important because most automatic methods for orthology prediction, like OrthoMCL, do not require proteins to be full length homologs.

2.6.3.2 Pairwise conservation of functional parameters

To measure the conservation of co-expression, first the correlation between the expression profiles of each human-human gene pair was calculated. The expression dataset used was a subset of pathologically normal human and mouse tissue samples from the Gene Logic BioExpress Database product, as mentioned above. This time we used all of the 115 categories to calculate the Pearson correlation coefficient for the human-human pairs, and we calculated the Pearson correlation coefficients for the mouse-mouse gene pairs using the 25 tissue categories in mouse. Co-expression is considered conserved when the studied human gene pair having a Pearson correlation coefficient above a certain threshold has an orthologous gene pair in mouse that has a Pearson correlation coefficient above the same threshold. This threshold was varied between 0.0 and 1.0 with an interval of 0.1. Co-expression can be used to predict protein function, specifically when it is conserved in evolution [10,15]. To test which of the ortholog sets can best be used to improve co-expression based function prediction, we also determined which protein pairs were active in the same process, using the GO database [16]. Two proteins were said to be active in the same process if they shared a 4th level element of the GO biological process tree, in which the root is the 0th level element and every subsequent branch is one level higher. Finally, the fraction of the total protein set sharing this 4th level element was calculated for the several thresholds, as a measure for the sensitivity and selectivity of the ortholog identification method for function prediction by conservation of co-expression. In this analysis, GO labels such as 'undefined' were discarded. The human-worm analysis was performed in a similar way, but with the use of expression data from Stuart and colleagues [38]. For calculating reliable correlation coefficients, we only used genes here that had expression data for at least 900 out of the 1,202 human tissue samples. In worm, we used all genes having expression data for at least 500 out of the 979 tissue samples.

The conservation of gene order is the second measure of pairwise conservation. Here we examined if two genes were adjacent to each other on the genome using EnsMart [30] version 19.3 [31] for the human-mouse analysis and EnsMart version 34 for the human-worm analysis. For each of the pairs where this was the case, we examined if the orthologs in mouse/worm were also adjacent on the genome. If so, the gene order was considered to be conserved for this gene pair. Because no varying threshold is needed (two genes are adjacent or not), this is more straight-forward than measuring the conservation of co-expression. The fraction of neighboring human genes of which the orthologs in mouse/worm are also neighbors is used as a measure for the accuracy of orthology prediction.

A third measure of pairwise conservation is the conservation of protein-protein interaction. The Database of Interacting Proteins (DIP) database [40] was used to determine the protein-protein interactions in human and mouse/worm. A protein-protein interaction is considered conserved when two interacting proteins in human have orthologs in mouse/worm that are interacting too. Again, the fraction of interacting human proteins of which the orthologs in mouse/worm are interacting too is considered to be a measure for the conservation of function.

2.6.4 Ortholog reference set

We defined a list of 'true ortholog pairs', for both human-mouse and human-worm, as a reference set. We chose the Hox cluster proteins and hemoglobins as a human-mouse reference set because of its well-studied evolution in vertebrates. We determined the homeobox orthologs using figure 1 from [41]. This resulted in 41 orthologous protein pairs, consisting of 31 human proteins and 35 mouse proteins. The hemoglobin orthologs were identified with the use of Lecomte et al. [42], resulting in nine pairs of four human and nine mouse proteins. For human-worm, we used the analysis on nuclear receptors performed by Gissendanner et al. [43], resulting in 29 orthologous pairs of 22 human proteins and 18 worm proteins. A second human-worm orthology analysis was performed on the family of toll-like receptors [44], which has only one member in worm but 10 members in human. The fifth and final protein family, the Sm and Sm-like proteins [45], was analyzed for both human-mouse and human-worm orthologs. For this family we found 13 human proteins and 17 mouse proteins in 17 orthologous pairs, together with 6 human proteins and 6 worm proteins in 6 pairs.

For each of these parts of our reference set and for each of the six ortholog identification methods, we determined how many of these orthologous pairs were covered, together with the number of false positives (pairs having only the human protein or the mouse/worm protein from a reference pair). Finally, to have a fair comparison between the several ortholog identification methods, we calculated the number of orthologous pairs divided by the average proteome size.

2.7 Additional data files

The following additional data are available with the online version of this paper (http://www.genomebiology.com/2006/7/4/R31). Additional data file 1 contains all end data used to create the figures. Additional data file 2 contains all of the protein pairs that are considered to be 'true orthologs' within our ortholog reference set, consisting of several protein families. The first column contains the name of the protein family, the second the human gene names and the third the mouse/worm gene names. The fourth column contains the corresponding human 'Protein World' entries, whereas the fifth column contains the mouse/worm entries. The last columns contain the orthologous protein pairs.

2.8 Acknowledgements

This work was supported financially by NV Organon and the Netherlands Organization for Scientific Research (NWO). The authors thank Gene Logic Inc. for the use of a subset of pathologically normal human and mouse tissue samples from the Gene Logic BioExpress Database product.

2.9 References

1. Fitch WM: Distinguishing homologous from analogous proteins. *Syst Zool* 1970, 19:99-113.
2. Li WH, Yang J, Gu X: Expression divergence between duplicate genes. *Trends Genet* 2005, 21:602-607.
3. Mirny LA, Gelfand MS: Using orthologous and paralogous proteins to identify specificity determining residues. *Genome Biol* 2002, 3:PREPRINT0002.
4. Chimpanzee sequencing whitepaper [http://genome.wustl.edu/ancillary/data/whitepapers/Pan_troglodytes_WP2.pdf]
5. Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000, 28:33-36.
6. Remm M, Storm CE, Sonnhammer EL: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001, 314:1041-1052.
7. Li L, Stoeckert CJ Jr, Roos DS: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003, 13:2178-2189.
8. Huynen MA, Bork P: Measuring genome evolution. *Proc Natl Acad Sci* USA 1998, 95:5849-5856.
9. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al.: The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003, 4:41.
10. van Noort V, Snel B, Huynen MA: Predicting gene function by conserved co-expression. *Trends Genet* 2003, 19:238-242.
11. Wright MW, Eyre TA, Lush MJ, Povey S, Bruford EA: HCOP: the HGNC comparison of orthology predictions search tool. *Mamm Genome* 2005, 16:827-828.
12. Zmasek CM, Eddy SR: RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 2002, 3:14.
13. Huynen MA, Snel B, van Noort V: Comparative genomics for reliable protein-function prediction from genomic data. *Trends Genet* 2004, 20:340-344.
14. Sjolander K: Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 2004, 20:170-179.

15. Stuart JM, Segal E, Koller D, Kim SK: A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003, 302:249-255.
16. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al.: The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004, 32 (Database):D258-261.
17. Fraser HB, Hirsh AE, Wall DP, Eisen MB: Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci* USA 2004, 101:9033-9038.
18. Protein World Webserver [http://www.cmbi.ru.nl/pw/]
19. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003, 31:365-370.
20. Pruitt KD, Tatusova T, Maglott DR: NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2005, 33 (Database):D501-504.
21. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: The Ensembl automatic gene annotation system. *Genome Res* 2004, 14:942-950.
22. Smith TF, Waterman MS: Identification of common molecular subsequences. *J Mol Biol* 1981, 147:195-197.
23. Comet JP, Aude JC, Glemet E, Risler JL, Henaut A, Slonimski PP, Codani JJ: Significance of Z-value statistics of Smith-Waterman scores for protein alignments. *Comput Chem* 1999, 23:317-331.
24. Brenner SE, Chothia C, Hubbard TJ: Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA* 1998, 95:6073-6078.
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403-410.
26. Pearson WR, Lipman DJ: Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988, 85:2444-2448.
27. Center for Molecular and Biomolecular Informatics [http://www.cmbi.ru.nl/]
28. InParanoid Program [http://inparanoid.cgb.ki.se/prog/inparanoid.tar.gz]
29. Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002, 30:1575-1584.
30. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: EnsMart: a generic system for fast and flexible access to biological data. *Genome Res* 2004, 14:160-169.
31. EnsMart v. 19.3 [ftp://ftp.ensembl.org/pub/current_mart/]
32. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS: OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 2006, 34(Database):D363-368.
33. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994, 22:4673-4680.

34. Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987, 4:406-425.
35. Supplementary Data: Orthology Comparison [http://www.cmbi.ru.nl/~timhulse/orthocomp/]
36. Gene Logic BioExpress Database Product [http://www.genelogic.com/genomics/bioexpress/]
37. Cote RA, Robboy S: Progress in medical information management. Systematized nomenclature of medicine (SNOMED). *JAMA* 1980, 243:756-762.
38. Supplementary Data for Stuart et al. [15] [http://cmgm.stanford.edu/~kimlab/multiplespecies/Data/]
39. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, et al.: The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res* 2003, 31:315-318.
40. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 2002, 30:303-305.
41. Pollard SL, Holland PW: Evidence for 14 homeobox gene clusters in human genome ancestry. *Curr Biol* 2000, 10:1059-1062.
42. Lecomte JT, Vuletich DA, Lesk AM: Structural divergence and distant relationships in proteins: evolution of the globins. *Curr Opin Struct Biol* 2005, 15:290-301.
43. Gissendanner CR, Crossgrove K, Kraus KA, Maina CV, Sluder AE: Expression and function of conserved nuclear receptor genes in Caenorhabditis elegans. *Dev Biol* 2004, 266:399-416.
44. Zheng L, Zhang L, Lin H, McIntosh MT, Malacrida AR: Toll-like receptors in invertebrate innate immunity. *Invertebrate Survival J* 2005, 2:105-113.
45. Salgado-Garrido J, Bragado-Nilsson E, Kandels-Lewis S, Seraphin B: Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. *EMBO J* 1999, 18:3451-3462.

Chapter 3

Testing statistical significance scores of sequence comparison methods with structure similarity

Tim Hulsen, Jacob de Vlieg, Jack A.M. Leunissen and Peter M.A. Groenen

BMC Bioinformatics 2006, **7** (1): 444

3.1 Abstract

3.1.1 Background

In the past years the Smith-Waterman sequence comparison algorithm has gained popularity due to improved implementations and rapidly increasing computing power. However, the quality and sensitivity of a database search is not only determined by the algorithm but also by the statistical significance testing for an alignment. The e-value is the most commonly used statistical validation method for sequence database searching. The CluSTr database and the Protein World database have been created using an alternative statistical significance test: a Z-score based on Monte-Carlo statistics. Several papers have described the superiority of the Z-score as compared to the e-value, using simulated data. We were interested if this could be validated when applied to existing, evolutionary related protein sequences.

3.1.2 Results

All experiments are performed on the ASTRAL SCOP database. The Smith-Waterman sequence comparison algorithm with both e-value and Z-score statistics is evaluated, using ROC, CVE and AP measures. The BLAST and FASTA algorithms are used as reference. We find that two out of three Smith-Waterman implementations with e-value are better at predicting structural similarities between proteins than the Smith-Waterman implementation with Z-score. SSEARCH especially has very high scores.

3.1.3 Conclusions

The compute intensive Z-score does not have a clear advantage over the e-value. The Smith-Waterman implementations give generally better results than their heuristic counterparts. We recommend using the SSEARCH algorithm combined with e-values for pairwise sequence comparisons.

3.2 Background

Sequence comparison is still one of the most important methodologies in the field of computational biology. It enables researchers to compare the sequences of genes or proteins with unknown functions to sequences of well-studied genes or proteins. However, due to a significant increase in whole genome sequencing projects, the amount of sequence data is nowadays very large and rapidly increasing. Therefore, pairwise comparison algorithms should not only be accurate and reliable but also fast. The Smith-Waterman algorithm [1] is one of the most advanced and sensitive pairwise sequence comparison algorithms currently available. However, it is theoretically about 50 times slower than other popular algorithms [2], such as FASTA [3] and BLAST [4]. All three algorithms generate local alignments, but the Smith-Waterman algorithm puts no constraints on the alignment it reports other than that it has a positive score in terms of the similarity table used to score the alignment. BLAST and FASTA put additional constraints on the alignments that they report in order to speed up their operation: only sequences above a certain similarity threshold are reported, the rest is used for the estimation of certain parameters used in the alignment calculation. Because of this Smith-Waterman is more sensitive than BLAST and FASTA. The Smith-Waterman algorithm finds the best matching regions in the same pair of sequences. However, BLAST and FASTA are still far more popular because of their speed and the addition of a statistical significance value, the Expect-value (or simply e-value), whereas the original Smith-Waterman implementation relies only on the SW-score without any further statistics. The newer Smith-Waterman implementations of Paracel [5], SSEARCH [6] and ParAlign [7] do include the e-value as a measure of statistical significance, which makes the Smith-Waterman algorithm more usable as the engine behind a similarity search tool. The e-value is far more useful than the SW-score, because it describes the number of hits one can expect to see by chance when searching a database of a certain size. An e-value threshold can be used easily to separate the 'interesting' results from the background noise. However, a more reliable statistical estimate is still needed [8]. The Z-score, based on Monte-Carlo statistics, was introduced by Doolittle [9] and implemented by Gene-IT [10] in its sequence comparison suite Biofacet [11]. The Z-score has been used in the creation of the sequence annotation databases CluSTr [12] and Protein World [13] and was used in orthology studies [14]. The Z-score has also been implemented in algorithms other than Smith-Waterman, such as FASTA [15]. It is calculated by performing a number (e.g., 100) of shuffling randomizations of both sequences that are compared, completed by an estimation of the SW score significance as compared to the original pairwise alignment. This makes the Z-score very useful for doing all-against-all pairwise sequence comparisons: Z-scores of different sequence pairs can be compared to each other, because they are only dependent on the sequences itself and not on the database size, which is one of the parameters used to calculate the e-value. However, this independency of the database size makes the Z-score unsuitable for determining the probability that an alignment has been obtained by chance. The randomizations make the Z-score calculation quite slow, but theoretically it is more sensitive and more selective than e-value statistics [16, 17]. Unfortunately, this has never been validated experimentally.

Some methods have been used to combine the sensitivity and selectivity of a sequence comparison algorithm into one single score [18]. Receiver operating characteristic (ROC) is a popular measure of search accuracy [19]. For a perfect search algorithm, all true positives for these queries should appear before any false positive in the ranked output list, which gives an ROC score of 1. If the first n items in the list are all false positives, the ROCn score is 0. Although researchers have devised many ways to merge ROC scores for a set of queries [20], one simple and popular method is to 'pool' search results so as to get an overall ROC score [21]. Another method to evaluate different methods is the errors per query (EPQ) criterion and the 'coverage versus error' plots [2]. EPQ is a selectivity indicator based on all-against-all comparisons, and coverage is a sensitivity measure. The assumption for EPQ is that the search algorithm can yield a 'normalized similarity score' rather than a length-dependent one, so that results from queries are comparable. Like ROC, the coverage versus error plot can give an overall performance comparison for search algorithms. A third method, the average precision (AP) criterion, is adopted from information retrieval research [22]. The method defines two values: the recall (true positives divided by the number of homologs) and the precision (true positives divided by the number of hits), which are plotted in a graph. The AP then is an approximate integral to calculate the area under this recall-precision curve.

These methods were used to compare several sequence comparison algorithms, but we use them to compare the e-value and Z-score statistics. Analyses of BLAST and FASTA are also included as reference material.

Here we show that two out of the three Smith-Waterman implementations with e-value statistics are more accurate than the Smith-Waterman implementation of Biofacet with Z-score statistics. Furthermore, the comparison of BLAST and FASTA with the four Smith-Waterman implementations shows that FASTA is a more reliable algorithm when using the ASTRAL SCOP structural classification as a benchmark. The Smith-Waterman implementation of Paracel even has lower scores than both BLAST and FASTA. SSEARCH, the Smith-Waterman implementation in the FASTA package, scores best.

3.3 Results

We used a non-redundant protein-domain sequence database derived from PDB as the target database. It is automatically generated using the ASTRAL system [23]. According to the structural classification of proteins (SCOP release 1.65), it includes 9498 sequences and 2326 families. True positives are those in the same family as the query sequence. SCOP as an independent and accurate source for evaluating database search methods has been used by other researchers [2, 24]. ASTRAL SCOP sets with different maximal percentage identity thresholds (10%, 20%, 25%, 30%, 35%, 40%, 50%, 70%, 90% and 95%) were downloaded from the ASTRAL SCOP website [25]. Their properties (number of families, number of members, etc.) are shown in table 1. Three different statistical measures were applied: receiver operating characteristic (ROC), coverage versus error (CVE) and mean average precision (AP). We compared six different pairwise sequence comparison algorithms, which are listed in table 2, together with the parameters used in this study.

Table 1. Properties of ASTRAL SCOP PDB sets

Maximal percentage indentity	Number of sequences	Number of families	Average family size	Size of largest family	Number of families having only 1 member	Number of families having more than 1 member
10%	3631	2250	1.614	25	1655	595
20%	3968	2297	1.727	29	1605	692
25%	4357	2313	1.884	32	1530	783
30%	4821	2320	2.078	39	1435	885
35%	5301	2322	2.283	46	1333	989
40%	5674	2322	2.444	47	1269	1053
50%	6442	2324	2.772	50	1178	1146
70%	7551	2325	3.248	127	1087	1238
90%	8759	2326	3.766	405	1023	1303
95%	9498	2326	4.083	479	977	1349

Table 2. Sequence comparison methods and parameters

Method	Abbreviation	Version	Matrix	Gap open penalty	Gap extension penalty	Number of randomizations
Paracel SW e-value	pc e	-	BLOSUM62	3*IS *	0.3*IS *	0
Biofacet SW Z-score	bf z	2.9.6	BLOSUM62	12	1	100

NCBI BLAST e-value	bl e	2.2.9	BLOSUM62	12	1	0
FASTA e-value	fa e	3.4t24	BLOSUM62	12	1	0
SSEARCH e-value	ss e	3.4t24	BLOSUM62	12	1	0
ParAlign SW e-value	pa e	4.0.0	BLOSUM62	12	1	0

* IS = average matrix identity score

3.3.1 Receiver operating characteristic

The mean ROC50 scores increase if more structurally identical proteins are included, for both the e-value and the Z-score measurements (figure 1). The ROC50 scores of the PDB010 set show a large difference between the several Smith-Waterman implementations: 0.19 for Paracel, 0.23 for Biofacet (with Z-score), 0.27 for ParAlign and 0.31 for SSEARCH. The advantage of ParAlign over Biofacet decreases with increasing inclusiveness of the ASTRAL SCOP set that is used. The ROC50 scores of the PDB095 set are 0.28 for Paracel, 0.35 for both ParAlign and Biofacet (with Z-score) and 0.46 for SSEARCH. SSEARCH scores best of all studied methods, regardless of which ASTRAL SCOP set is used. The reference methods FASTA and BLAST give quite different results: FASTA is a good second and BLAST has scores similar to Paracel and Biofacet.

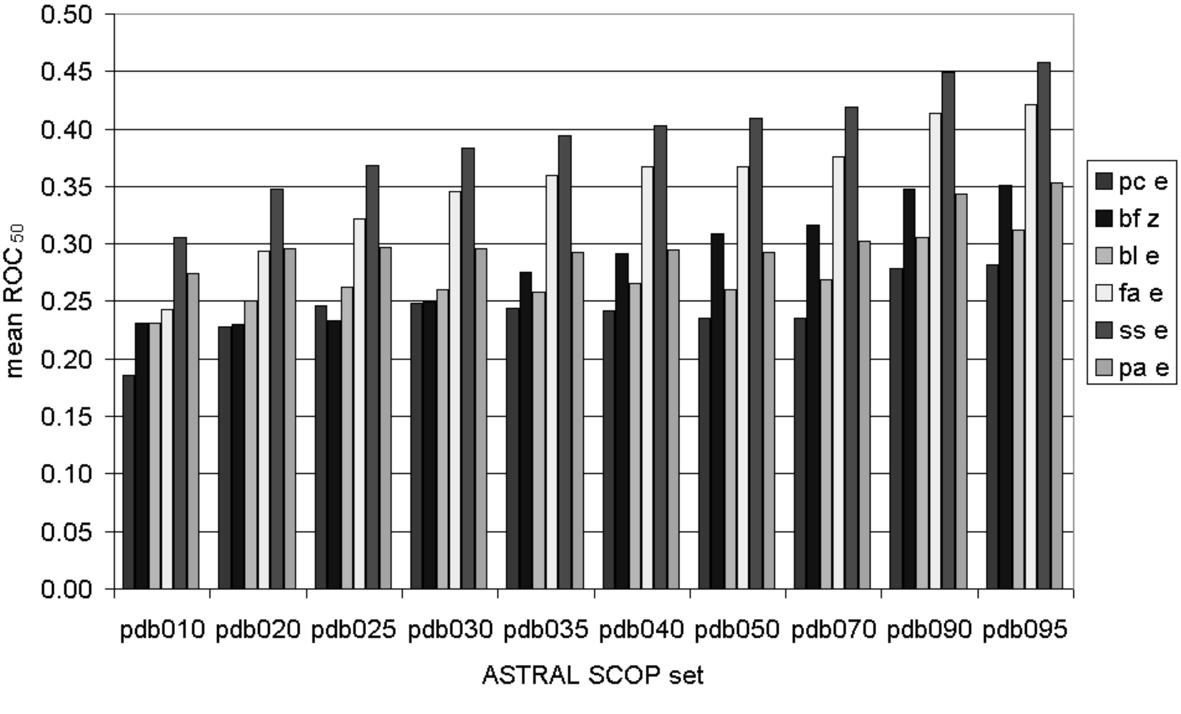
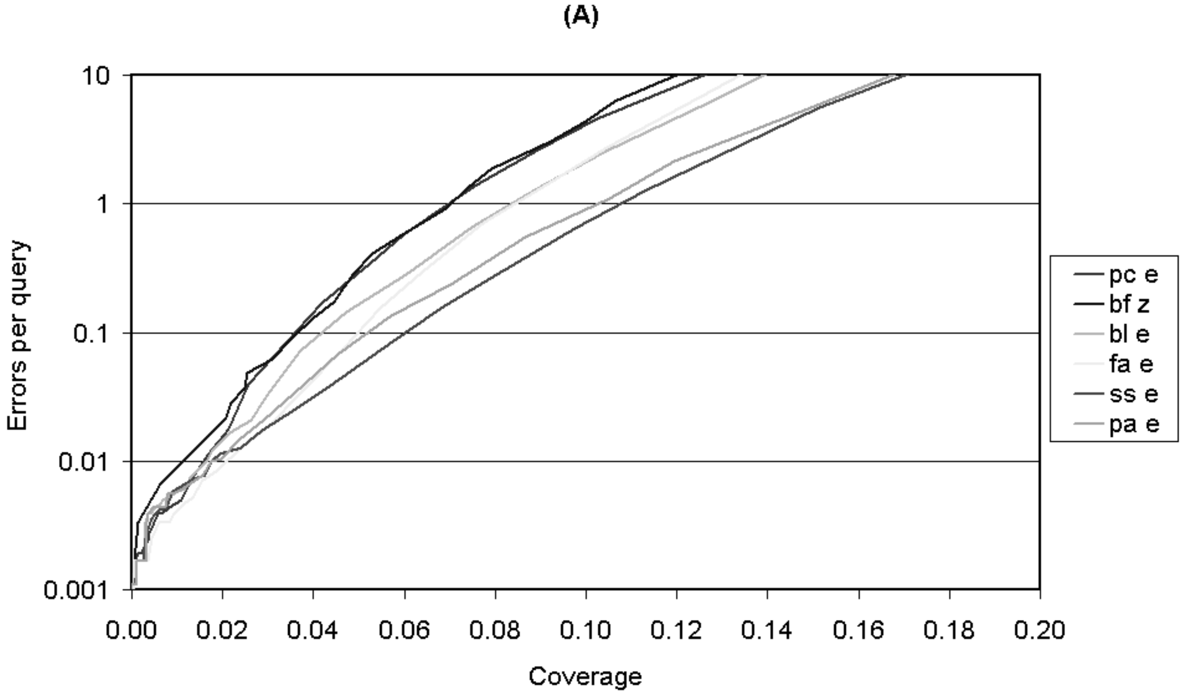


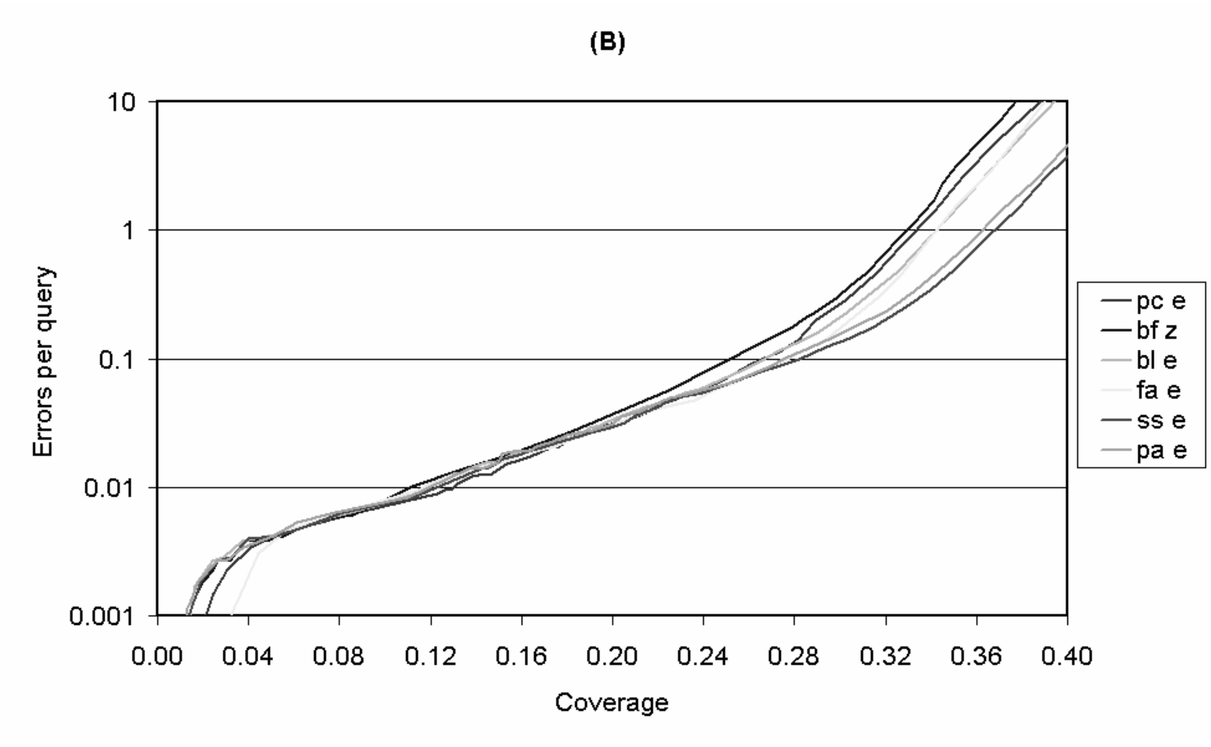
Figure 1. The mean Receiver Operating Characteristic scores for ten different ASTRAL SCOP sets

The maximal structural identity percentage of each set increases from the left to the right, from 10% to 95%. Red bars: mean ROC$_{50}$ scores calculated using the Paracel Smith-Waterman algorithm. Blue bars: mean ROC$_{50}$ scores calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Green bars: mean ROC$_{50}$ scores calculated using the BLAST algorithm. Yellow bars: mean ROC$_{50}$ scores calculated using the FASTA algorithm. Purple bars: mean ROC$_{50}$ scores calculated using the SSEARCH algorithm. Orange bars: mean ROC$_{50}$ scores calculated using the ParAlign Smith-Waterman algorithm. *Color version on page 153.*

3.3.2 Coverage versus error

This method differs from the ROC analysis on one crucial point: instead of looking at the first 100 hits, we varied the threshold at which a hit was seen as a positive. Hence the results are somewhat dissimilar: the differences between the several algorithms in the coverage versus error plots (figure 2) are not as obvious as they are in the ROC50 graph (figure 1). Figure 2a shows the coverage versus error plot for the smallest ASTRAL SCOP set (PDB010), figure 2b shows the plot for the largest ASTRAL SCOP set (PDB095) and figure 2c shows the plot for the intermediate set PDB035. An ideal algorithm would have a very high coverage but not many errors per query, which places it in the lower right corner of the graph. SSEARCH has the best scores when using the PDB010 set, followed by ParAlign and FASTA, with the latter scoring best in the lowest-coverage range (<0.02). Biofacet with Z-score has the lowest scores. The PDB095 plot shows some differences between the low-coverage range (<0.25) and the high-coverage range (>0.50). In the low coverage range, FASTA and Paracel have the highest scores, whereas SSEARCH and ParAlign have the highest scores in the low-coverage range. It should be noted that the high-coverage range might statistically be more reliable because of the larger number of hits. The PDB035 set gives similar results.





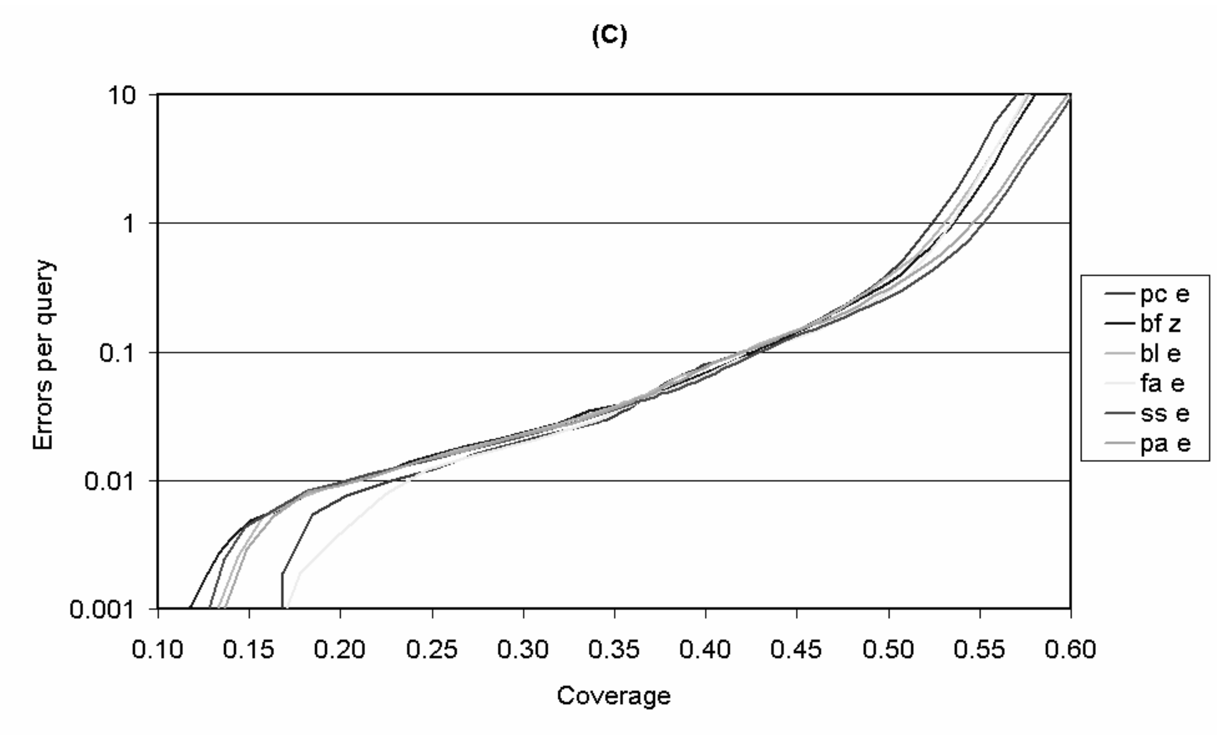


Figure 2. (a) Coverage versus error plot for the ASTRAL SCOP PDB010 set. **(b)** Coverage versus error plot for the ASTRAL SCOP PDB035 set. **(c)** Coverage versus error plot for the ASTRAL SCOP PDB095 set.

Red line: calculated using the Paracel Smith-Waterman algorithm. Blue line: calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Green line: calculated using the BLAST algorithm. Yellow line: calculated using the FASTA algorithm. Purple line: calculated using the SSEARCH algorithm. Orange line: calculated using the ParAlign Smith-Waterman algorithm. *Color version on page 154-155.*

3.3.3 Average precision

The average precision graph (figure 3) shows some minor differences from the ROC50 graph (figure 1): for the PDB020, PDB025 and PDB030 set, Paracel (e-value) scores better than Biofacet (Z-score). However, the advantage of the Biofacet Smith-Waterman with Z-score increases from that point on (PDB035, Paracel: 0.16, Biofacet: 0.17) to the right side (PDB095, Paracel: 0.19, Biofacet: 0.24). The Z-score seems to score better when more similar proteins are compared. Once more, SSEARCH has the highest scores for all structural identity percentages, with FASTA as the second best.

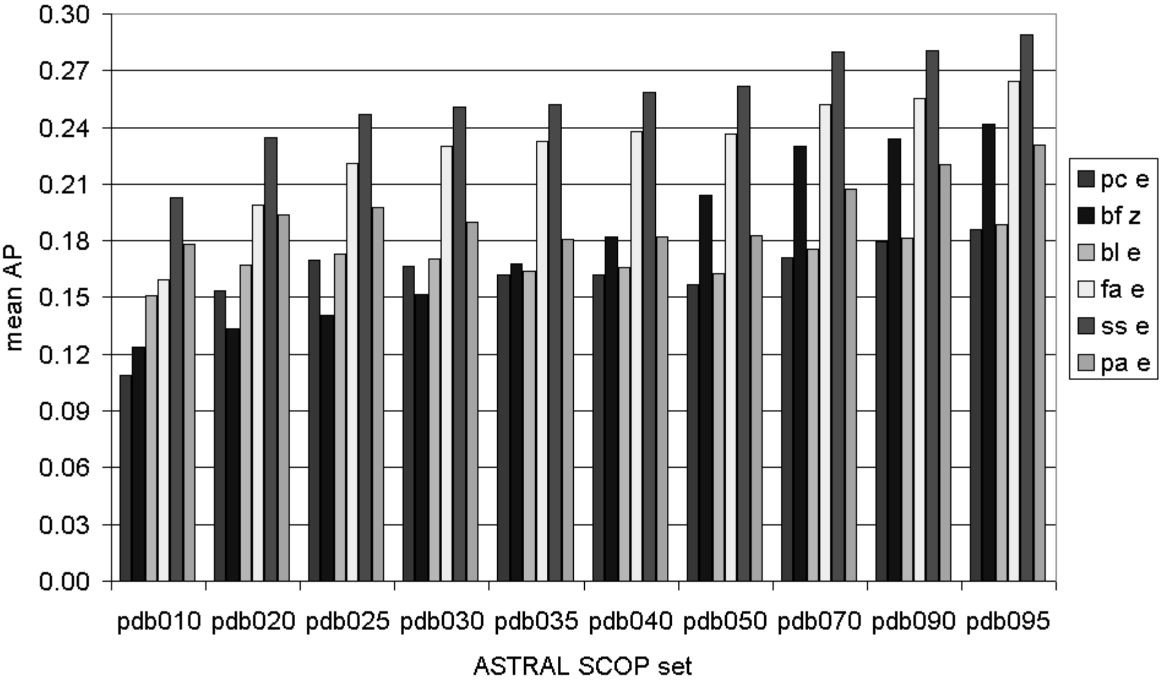


Figure 3. The average precision values for ten different ASTRAL SCOP sets
The maximal structural identity percentage of each set increases from the left to the right, from 10% to 95%. Red bars: mean AP values calculated using the Paracel Smith-Waterman algorithm. Blue bars: mean AP values calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Green bars: mean AP values calculated using the BLAST algorithm. Yellow bars: mean AP values calculated using the FASTA algorithm. Purple bars: mean AP values calculated using the SSEARCH algorithm. Orange bars: mean AP values calculated using the ParAlign Smith-Waterman algorithm. *Color version on page 156.*

3.3.4 Case studies

We included two examples of our statistical analysis, which show how the ROC and mean AP measures differ from each other and how results can be different for each studied protein. We choose two well-studied proteins: enoyl-ACP reductase and the progesterone receptor, the first from a prokaryote (*E. coli*) and the second from a eukaryote (*H. sapiens*). Both case studies were done using the PDB095 set, which is the most complete ASTRAL SCOP PDB set used in our study.

3.3.4.1 Bacterial enoyl-ACP reductase

Table 3 shows the results of our analysis of the ASTRAL SCOP entry of *E. coli* enoyl-ACP reductase chain A, d1qg6a_, using the PDB095 set. One way of testing the reliability of a sequence comparison method is by looking at the first false positive (FFP) in the list of top 100 hits (Table S.1 [see Additional file 1]). The c.2.1.2 structural family has 46 members within the PDB095 set, so the perfect sequence comparison algorithm would return its first false positive at the 46th hit (the hit containing the query protein is discarded). For the Paracel Smith-Waterman implementation, this is already the twenty-first hit. Four algorithms score best with the first false positive at 24th place. A second testing method is counting the total number of true positives (NTP), of which the perfect algorithm would return all 45. BLAST has the highest score here: 27 out of the top 100 hits are true positives. FASTA and Paracel are at the second place with 25 true positives. Biofacet has the lowest score: only 23 true positives. Note that differences are very small, which is a reason to look at the ROC and mean AP scores. FASTA and SSEARCH have both the highest ROC50 scores and the highest mean APs. ParAlign and BLAST are third and fourth, followed by Paracel and Biofacet. The ROC and mean AP scores give a clearer view of the differences between the algorithms than the FFP or NTP scores, because they take into account the ranking of all hits instead of just the first false positive or just the true positives.

Table 3. Scores for bacterial enoyl-ACP reductase

***E. coli* enoyl-ACP reductase**	**pc e**	**bf z**	**bl e**	**fa e**	**ss e**	**pa e**
ROC score	0.156	0.124	0.250	0.367	0.338	0.229
MAP score	0.212	0.161	0.264	0.374	0.343	0.234
First False Polsitive (FFP)	21	24	24	22	24	24
Number of True Positives (NTP)	25	23	27	25	24	24

3.3.4.2 Human progesterone receptor

Table 4 shows our analysis of ASTRAL SCOP entry d1a28a_, using again the PDB095 set. The structural family a.123.1.1 has 29 members, so the perfect algorithm should have the first false positive at the 29th hit. Surprisingly, BLAST scores best here with its first false positive at the 25th hit (Table S.2 [see Additional file 1]), although the differences are quite small. BLAST is, together with Biofacet, the only algorithm that does not have all the 28 family members of d1a28a_ in its top 100 list; d1n83a_ is missing here. The ROC50 and mean AP analysis of d1a28a_ shows again that SSEARCH and FASTA give the best results. Paracel and Biofacet have the lowest scores once more. The differences are not large enough to put any definite conclusions to the results of this example, but by combining all ROC and mean AP scores for all ASTRAL SCOP entries, we created a reliable comparison between all sequence comparison methods.

Table 4. Scores for the human progesterone receptor

***H. sapiens* progesterone receceptor**	**pc e**	**bf z**	**bl e**	**fa e**	**ss e**	**pa e**
ROC score	0.402	0.437	0.513	0.745	0.762	0.573
MAP score	0.504	0.503	0.548	0.727	0.745	0.586
First False Positive (FFP)	22	18	25	23	23	23
Number of True Positives (NTP)	28	27	27	28	28	28

3.3.4.3 Timing

Table 5 shows the time that each of the six algorithms needs to perform an all-against-all sequence comparison of the ASTRAL SCOP PDB095 set. The BLAST algorithm is clearly the fastest, followed by the other heuristic algorithm FASTA. Of the Smith-Waterman algorithms, ParAlign is by far the fastest. The Biofacet algorithm needs much time to calculate 2 x 100 randomizations and is therefore the slowest sequence comparison algorithm.

Table 5. Time for all-against-all sequence comparison of the ASTRAL SCOP PDB095 set.

Method	Time
Paracel SW e-value	3 hours *
Biofacet SW Z-score	multiple days
NCBI BLAST e-value	15 minutes
FASTA e-value	40 minutes
SSEARCH e-value	5 hours, 49 minutes
ParAlign SW e-value	47 minutes

* estimation because of unavailability Paracel system

3.4 Discussion

The theoretical advantage of the Z-score over the e-value appears to be rejected by our results. Our results show that the e-value calculation gives an advantage over the computationally intensive Z-score, at least when looking only at the results from the Smith-Waterman algorithm. Some caution should be taken however, drawing any definite conclusions. First, the Z-score was designed to make a distinction between significant hits and non-significant hits that have high SW scores. It might have an advantage over the e-value when applied to the top hits only, but might have less advantage for the hits with lower SW scores. This idea is supported by the fact that the Z-score is better at scoring high-similarity sequence pairs. This is also reflected in the different ROC and AP scores for the PDB010 set and the PDB095 set: the difference between Z-score and e-value increases when structurally more similar protein pairs are being included. Second, the Z-score can differ for each run, because of its different randomizations [17]. The standard deviation of the Z-score increases almost proportionally with the Z-score itself, i.e. for higher Z-scores the variance will be larger [16]. However, the Z-score increases its precision when more randomizations are calculated (2 x 100 in this study). Third, the PDB set is somewhat biased: it only contains crystallized proteins, and it contains no hypothetical proteins and membrane proteins. The crystallized proteins in the PDB are on average smaller than proteins included in large sequence databases such as the UniProt [26] database (figure 4), whereas the amino acid distribution is approximately the same for these databases (figure 5).

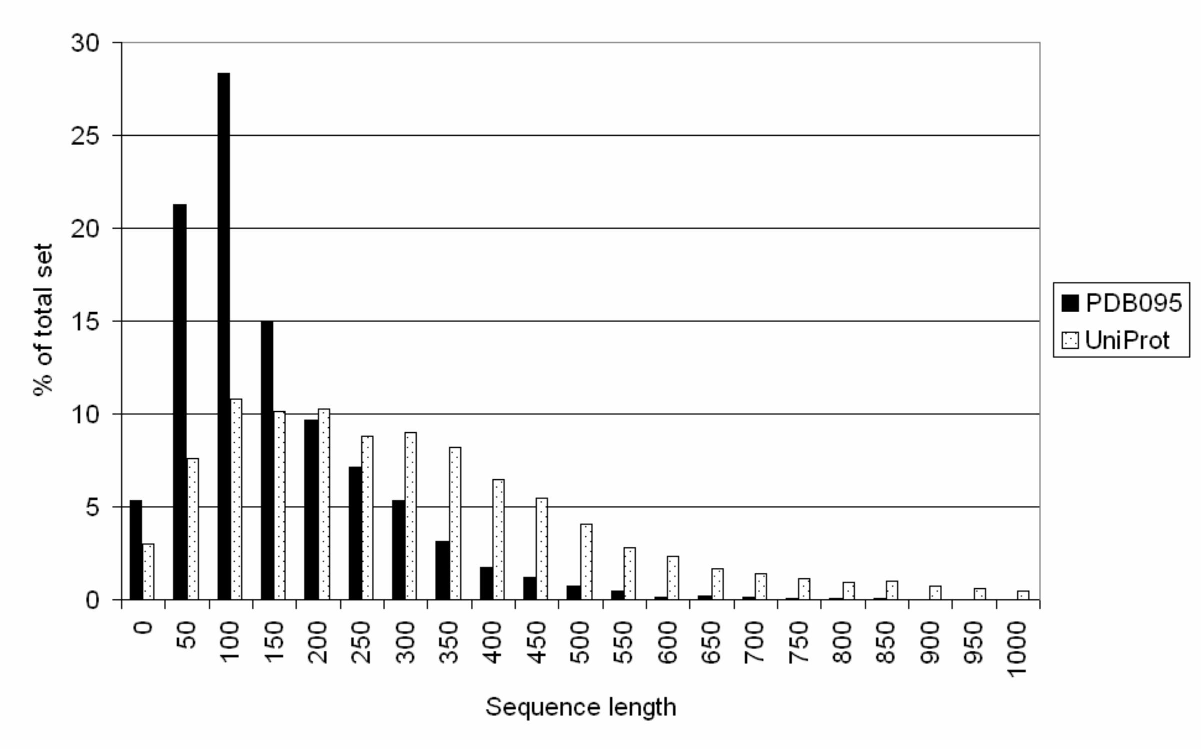


Figure 4. Sequence Length Distribution between PDB095 and UniProt

The sequence length increases from the left to the right. The vertical axis shows the number of proteins having that length, as a percentage of the total set. Black bars: PDB095 set. Dotted bars: UniProt set.

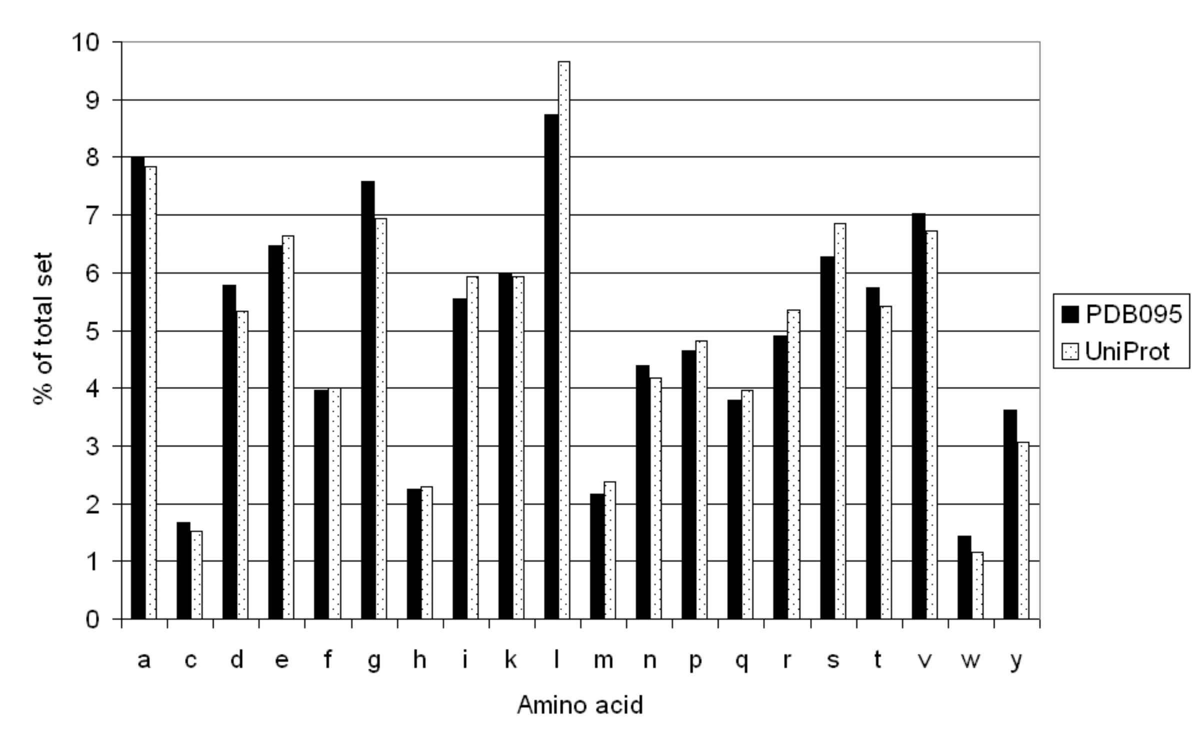


Figure 5. Amino Acid Distribution between PDB095 and UniProt

The 20 amino acids are displayed on the horizontal axis and their occurrence, as percentage of the total, is shown on the vertical axis. Black bars: PDB095 set. Dotted bars: UniProt set.

Figure 6 shows that the bias in sequence length is not the reason for the difference in scores: if we only look at proteins with a sequence length of 500 or more, the scores are similar. Other studies have shown that FASTA performs better than BLAST [18, 27], but these did not include several Smith-Waterman implementations. The SSEARCH algorithm, an implementation of Smith-Waterman, was analyzed in these studies, but this algorithm differs from other Smith-Waterman algorithms used in this study due to the use of length regression statistics [7, 28]. A difference can also be found by comparing the SW scores of Biofacet, ParAlign and SSEARCH: Biofacet and ParAlign have the same SW scores, but the SSEARCH SW scores are different. We calculated the ROC50 and mean AP for these three SW scores and found that the SSEARCH SW scores gives slightly worse results than the other two SW scores (figure 7). Another problem is that protein sequences within a certain ASTRAL SCOP family usually have equivalent lengths, since the ASTRAL SCOP database consists of protein domains and not of whole proteins. Results might vary when whole proteins, with different lengths, are studied. Unfortunately, the composition of the ASTRAL SCOP database does not allow us to confirm this statement.

Finally, we would like to stress that the results from the CVE analysis might be more reliable than those from the ROC and mean AP analyses. ROC and mean AP make use of a ranking system based on the e-value or Z-score, instead of looking at the e-value or Z-score directly. This means that in some cases, especially the smaller protein families, a large number of very low-scoring hits (e.g. e>100 or Z<3) is still used for the calculation of the scores. This is not the case for the CVE plots, because we varied the e-value and Z-score thresholds above which a hit is seen as a true positive, instead of relying on a ranking system. However, because the results from the CVE plots are similar to the results from the ROC and mean AP graphs, the use of a ranking system does not seem to give a large disadvantage.

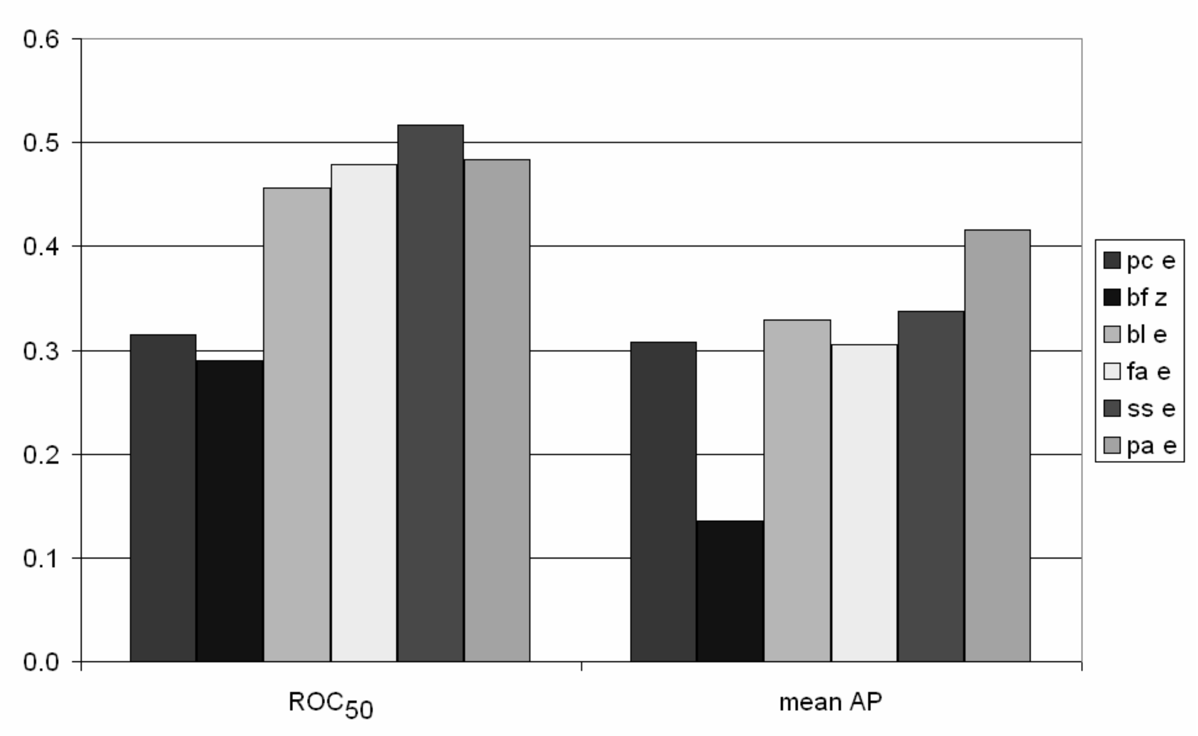


Figure 6. ROC$_{50}$ and mean AP values for proteins larger than 500 aa

The ROC_{50} scores are shown at the left half, the mean AP values on the right half. Red bars: calculated using the Paracel Smith-Waterman algorithm. Blue bars: calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Green bars: calculated using the BLAST algorithm. Yellow bars: calculated using the FASTA algorithm. Purple bars: calculated using the SSEARCH algorithm. Orange bars: calculated using the ParAlign Smith-Waterman algorithm. *Color version on page 157.*

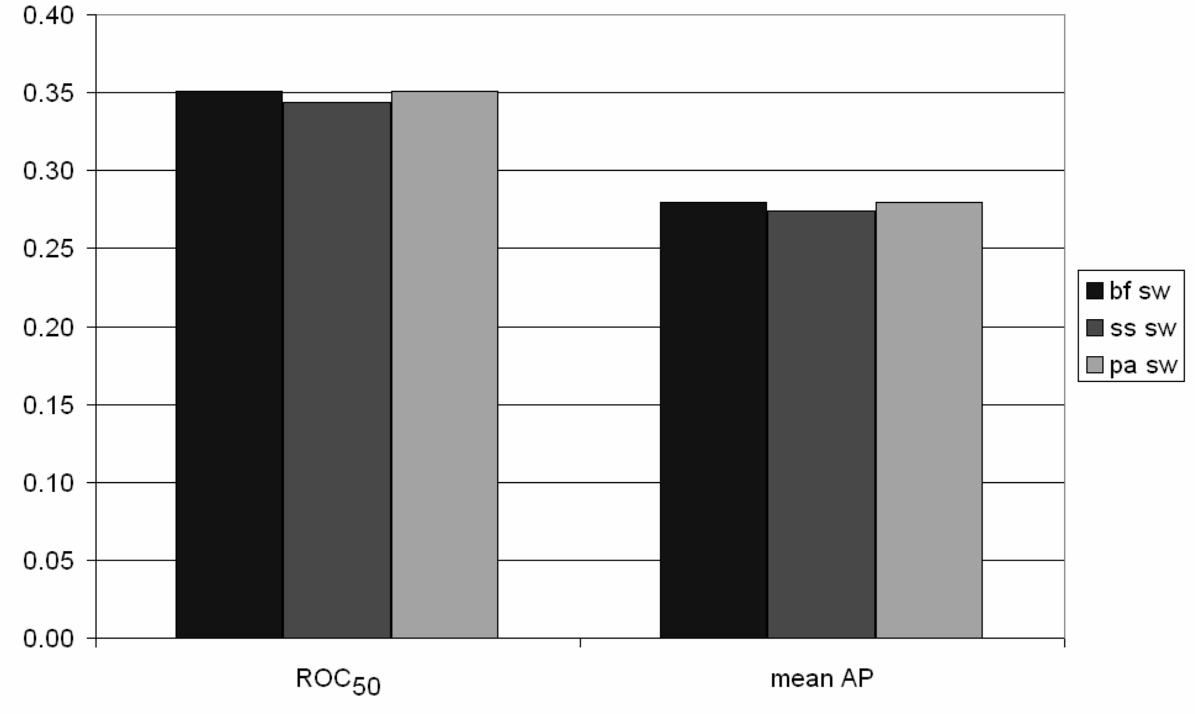


Figure 7. ROC_{50} and mean AP values for the SW scores of three different SW algorithms.
The ROC_{50} scores are shown at the left half, the mean AP values on the right half. Blue bars: calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Purple bars: calculated using the SSEARCH algorithm. Orange bars: calculated using the ParAlign Smith-Waterman algorithm. *Color version on page 157.*

3.5 Conclusions

For a complete analysis we need a less biased database, having a wide range of proteins classified by structure similarity. Until such a database is available, it will be difficult to pinpoint the reasons for the different results between FASTA, BLAST and Smith-Waterman, and the theoretical advantages of the Z-score. Regardless of all these theoretical assumptions, the computational disadvantage of the Z-score is smaller for larger databases. Z-scores do not have to be recalculated when sequences are added to the database, in contrast to e-values, which are dependent on database size. For very large databases containing all-against-all comparisons, this is an important advantage of the Z-score. Although recalculating the e-values does not take much time when the alignments and SW scores are already available, this may cause a change in research results that were obtained earlier. Despite these considerations, we recommend using SSEARCH with e-value statistics for pairwise sequence comparisons.

3.6 Methods

3.6.1 Sequence comparisons

For the Smith-Waterman e-value calculation, the ASTRAL SCOP files were loaded onto the Paracel file system as protein databases and subsequently used as queries against these databases: the set with 10% maximal identity (PDB010) against itself, the set with 20% maximal identity (PDB020) against itself, etc. The matrix used for all sequence comparisons was the BLOSUM62 matrix [29]. This is the default scoring matrix for most alignment programs. For all sequence comparisons in this article, the gap open penalty was set to 12 and the gap extension penalty was set to 1. These are the averages of the default penalties over the six studied methods. Both the matrix and gap penalties used are suited for comparing protein sets with a broad spectrum of evolutionary distances, like the PDB set [30, 31]. Per query sequence, the best 100 hits were kept [see section Data availability], discarding the match of each query sequence with itself.

3.6.2 Receiver operating characteristic calculation

For each query, the 100 best hits were marked as true positives or false positives, i.e. the hit being in the same or in a different SCOP family than the query. For each of the first 50 false positives that were found, the number of true positives with a higher similarity score was calculated. The sum of all of these numbers was then divided by the number of false positives (50), and finally divided by the total number of possible true positives in the database (i.e. the total number of members in the SCOP family minus 1), giving an ROC50 score for each query sequence. The average of these ROC50 scores gives the final ROC score for that specific statistical value and that specific ASTRAL SCOP set. Mean ROC50 scores were calculated for all ten different ASTRAL SCOP sets.

3.6.3. Coverage versus error calculation

Instead of taking the first 100 hits for each query, like in the ROC analysis, we varied the threshold at which a certain hit was seen as a positive. For the e-value analysis, we created a list of 49 thresholds in the range of 10-50 to 100. For Z-score, we created a list of 58 thresholds in the range of 0 to 100. Then, for each threshold, two parameters were measured: the coverage and the errors per query (EPQ). The coverage is the number of true hits divided by the total number of sequence pairs that are in the same SCOP family, for that specific ASTRAL SCOP set. The EPQ is the number of false hits divided by the number of queries. We used the most inclusive ASTRAL SCOP set (PDB095), the least inclusive set (PDB010) and an intermediate set (PDB035) to create the coverage versus error plots.

3.6.4 Average precision calculation

For the calculation of the average precision (AP), the 100 best hits per query were marked again as either true positives or false positives. Subsequently for each true positive found by the search algorithm, the true positive rank of this hit (i.e. the number of true positives with a higher score + 1) was divided by the positive rank (i.e.

the number of hits with a higher score + 1). These numbers were all added up and then divided by the total number of hits (i.e. 100), giving one AP value per query. The mean AP is the average of all these APs. Mean APs were calculated for all ten different ASTRAL SCOP sets.

3.6.5 Bacterial enoyl-ACP reductase

The ASTRAL SCOP entry for *E. coli* enoyl-ACP reductase chain A, d1qg6a_, was picked as an example for our methodology. The 100 best hits of this entry on the PDB095 set were calculated using each of the six algorithms and sorted by ascending e-value and descending Z-score. Then they were marked as either true positives or false positives, depending on if the hit was in the same structural family (c.2.1.2) or not. Furthermore, the ROC50 scores and mean APs were calculated.

3.6.6 Human progesterone receptor

A second example is the analysis of d1a28a_, the *H. sapiens* progesterone receptor chain A. Once more, the 100 best hits of this entry on the PDB095 set were calculated using each of the six algorithms and sorted by ascending e-value and descending Z-score. These hits were marked as either true positives or false positives, depending on if the hit was in the same structural family (a.123.1.1) or not. Finally, the mean AP and ROC50 scores were calculated.

3.6.7 Timing

We measured the speed of the sequence comparison algorithms, by doing an all-against-all comparison of the ASTRAL SCOP PDB095 set and using the ‘time’ command provided by UNIX. All calculations were performed on the same machine, except for the Paracel calculation which could only be performed on the Paracel machine. The Paracel calculation time had to be estimated because of the unaivailability of the Paracel machine at the time of performing this analysis.

3.7 Acknowledgements

This work was supported financially by NV Organon and the Netherlands Organization for Scientific Research (NWO). The authors like to thank Scott Lusher for critically reading this manuscript.

3.8 Data availability

All raw sequence comparison output files (containing the top 100 hits per query sequence) are available through our website [32]. The top 100 hits for the two case studies of the bacterial enoyl-ACP reductase (i.e. Table S.1) and the human progesterone receptor (i.e. Table S.2) can be found there as well.

3.9 References

1. Smith TF, Waterman MS: Identification of common molecular subsequences. *J Mol Biol* 1981, 147(1):195-197.
2. Brenner SE, Chothia C, Hubbard TJ: Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A* 1998, 95(11):6073-6078.
3. Pearson WR, Lipman DJ: Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 1988, 85(8):2444-2448.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25(17):3389-3402.
5. Paracel [http://www.paracel.com]
6. Pearson WR: Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 1991, 11(3):635-650.
7. Rognes T: ParAlign: a parallel sequence alignment algorithm for rapid and sensitive database searches. *Nucleic Acids Res* 2001, 29(7):1647-1652.
8. Pearson WR, Sierk ML: The limits of protein sequence comparison? *Curr Opin Struct Biol* 2005.
9. Doolittle RF: Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. Mill Valley California: University Science Books; 1986.
10. Gene-IT [http://www.gene-it.com]
11. Codani JJ, Comet JP, Aude JC, Glémet E, Wozniak A, Risler JL, Hénaut A, Slonimski PP: Automatic Analysis of Large-Scale Pairwise Alignments of Protein Sequences. *Methods in Microbiology* 1999, 28:229-244.
12. Kriventseva EV, Servant F, Apweiler R: Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Res* 2003, 31(1):388-389.
13. Protein World [http://www.cmbi.ru.nl/pw/]
14. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006, 7(4):R31.
15. Booth HS, Maindonald JH, Wilson SR, Gready JE: An efficient Z-score algorithm for assessing sequence alignments. *J Comput Biol* 2004, 11(4):616-625.
16. Comet JP, Aude JC, Glemet E, Risler JL, Henaut A, Slonimski PP, Codani JJ: Significance of Z-value statistics of Smith-Waterman scores for protein alignments. *Comput Chem* 1999, 23(3-4):317-331.
17. Bastien O, Aude JC, Roy S, Marechal E: Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics. *Bioinformatics* 2004, 20(4):534-537.
18. Chen Z: Assessing sequence comparison methods with the average precision criterion. *Bioinformatics* 2003, 19(18):2456-2460.
19. Gribskov M, Robinson NL: Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Compu Chem* 1996, 20:25-33.
20. Kester AD, Buntinx F: Meta-analysis of ROC curves. *Med Decis Making* 2000, 20(4):430-439.

21. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* 2001, 29(14):2994-3005.
22. Salton G: Developments in automatic text retrieval. *Science* 1991, 253:974-980.
23. Brenner SE, Koehl P, Levitt M: The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* 2000, 28(1):254-256.
24. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* 1998, 284(4):1201-1210.
25. ASTRAL SCOP release 1.65 [http://astral.berkeley.edu/scopseq-1.65.html]
26. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M et al: UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 2004, 32(Database issue):D115-119.
27. Agarwal P, States DJ: Comparative accuracy of methods for protein sequence similarity search. *Bioinformatics* 1998, 14(1):40-47.
28. Pearson WR: Comparison of methods for searching protein sequence databases. *Protein Sci* 1995, 4(6):1145-1160.
29. Henikoff S, Henikoff JG: Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992, 89(22):10915-10919.
30. Reese JT, Pearson WR: Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics* 2002, 18(11):1500-1507.
31. Price GA, Crooks GE, Green RE, Brenner SE: Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap. *Bioinformatics* 2005, 21(20):3824-3831.
32. Supplementary data [http://www.cmbi.ru.nl/~timhulse/ezcomp/]

Chapter 4

PhyloPat: phylogenetic pattern analysis of eukaryotic genes

Tim Hulsen, Jacob de Vlieg and Peter M.A. Groenen

BMC Bioinformatics 2006, **7** (1): 398

4.1 Abstract

4.1.1 Background

Background: Phylogenetic patterns show the presence or absence of certain genes or proteins in a set of species. They can also be used to determine sets of genes or proteins that occur only in certain evolutionary branches. Phylogenetic patterns analysis has routinely been applied to protein databases such as COG and OrthoMCL, but not upon gene databases. Here we present a tool named PhyloPat which allows the complete Ensembl gene database to be queried using phylogenetic patterns.

4.1.2 Description

PhyloPat is an easy-to-use webserver, which can be used to query the orthologies of all complete genomes within the EnsMart database using phylogenetic patterns. This enables the determination of sets of genes that occur only in certain evolutionary branches or even single species. We found in total 446,825 genes and 3,164,088 orthologous relationships within the EnsMart v40 database. We used a single linkage clustering algorithm to create 147,922 phylogenetic lineages, using every one of the orthologies provided by Ensembl. PhyloPat provides the possibility of querying with either binary phylogenetic patterns (created by checkboxes) or regular expressions. Specific branches of a phylogenetic tree of the 21 included species can be selected to create a branch-specific phylogenetic pattern. Users can also input a list of Ensembl or EMBL IDs to check which phylogenetic lineage any gene belongs to. The output can be saved in HTML, Excel or plain text format for further analysis. A link to the FatiGO web interface has been incorporated in the HTML output, creating easy access to functional information. Finally, lists of omnipresent, polypresent and oligopresent genes have been included.

4.1.3 Conclusions

PhyloPat is the first tool to combine complete genome information with phylogenetic pattern querying. Since we used the orthologies generated by the accurate pipeline of Ensembl, the obtained phylogenetic lineages are reliable. The completeness and reliability of these phylogenetic lineages will further increase with the addition of newly found orthologous relationships within each new Ensembl release.

4.2 Background

Phylogenetic patterns show the presence or absence of certain genes or proteins in a set of species. These patterns can be used to determine sets of genes or proteins that (COG) [1] which included a Phylogenetic Patterns Search (PPS) on its web interface. This phylogenetic pattern tool was further enhanced with the Extended Phylogenetic Patterns Search (EPPS) [2] tool, providing the possibility of querying the phylogenetic patterns of the COG protein database using regular expressions. The newest release of the OrthoMCL database, OrthoMCL-DB [3], also offers this possibility. However, suchs tool have only been available for querying proteins, and not for genes. The advantage of looking at gene families instead of protein families, is that the view on expansions and deletions is not distorted by any alternative transcripts and splice forms. The PhIGs [4], Hogenom [5] and TreeFam [6] databases all offer phylogenetic clustering of genes, but do not have the functionality of phylogenetic patterns. Here we introduce a web tool named PhyloPat that creates the possibility of querying all complete genomes of the highly reliable Ensembl [7] database using any phylogenetic pattern.

4.3 Construction & content

We generated a set of phylogenetic lineages containing all of the genes in Ensembl [7] that have orthologs in other species according to the EnsMart [8] database. This set covers all of the 21 (eukaryotic) species available in EnsMart version 40 (pre-versions and low coverage genomes not taken into account). We collected the complete set of orthologies between these species: 420 species pairs, 446,825 genes and 3,164,088 orthologous relationships. These orthologies consist of 2,000,706 one-to-one, 795,723 one-to-many and 367,659 many-to-many relationships, created by the very extensive orthology prediction pipeline [9] from Ensembl. This pipeline starts with the collection of a number of Best Reciprocal Hits (BRH, proven to be accurate [10]) and Best Score Ratio (BSR) values from a WUBlastp/Smith-Waterman whole-genome comparison. These are used to create a graph of gene relations, followed by a clustering step. These clusters are then applied to build a multiple alignment using MUSCLE [11] and a phylogenetic tree using PHYML [12]. Finally, the gene tree is reconciled with the species tree using RAP [5]. From each reconciled gene tree, the above mentioned orthologous relationships are inferred. After the collection of all orthologous pairs, we generated phylogenetic lineages using a single linkage algorithm. First, we determined the evolutionary order of the studied species using the NCBI Taxonomy [13] database. The phylogenetic tree of these species, together with some phylogenetic branch names, can be seen in figure 1. Second, we used this phylogenetic tree as a starting point for building our phylogenetic lineages. For each gene in the first species (*S. cerevisiae*), we looked for orthologs in the other 20 species. All orthologs were added to the phylogenetic lineage, and in the next round were checked for orthologs themselves, until no more orthologies were found for any of the genes. This process was repeated for all genes in all 21 species that were not yet connected to any phylogenetic lineage yet. The complete phylogenetic lineage determination generated 147,922 lineages. Please note that the phylogenetic order that we have determined here does not affect the construction of the phylogenetic lineages in any way: changing the order only influences the numbering of the phylogenetic lineages but not the contents of the lineages. This is due to our clustering method, in which each orthologous relationship is treated symmetrically. Figure 2 shows the database scheme: the phylogenetic lineages and some extra information have been stored in four tables, optimized for fast querying.

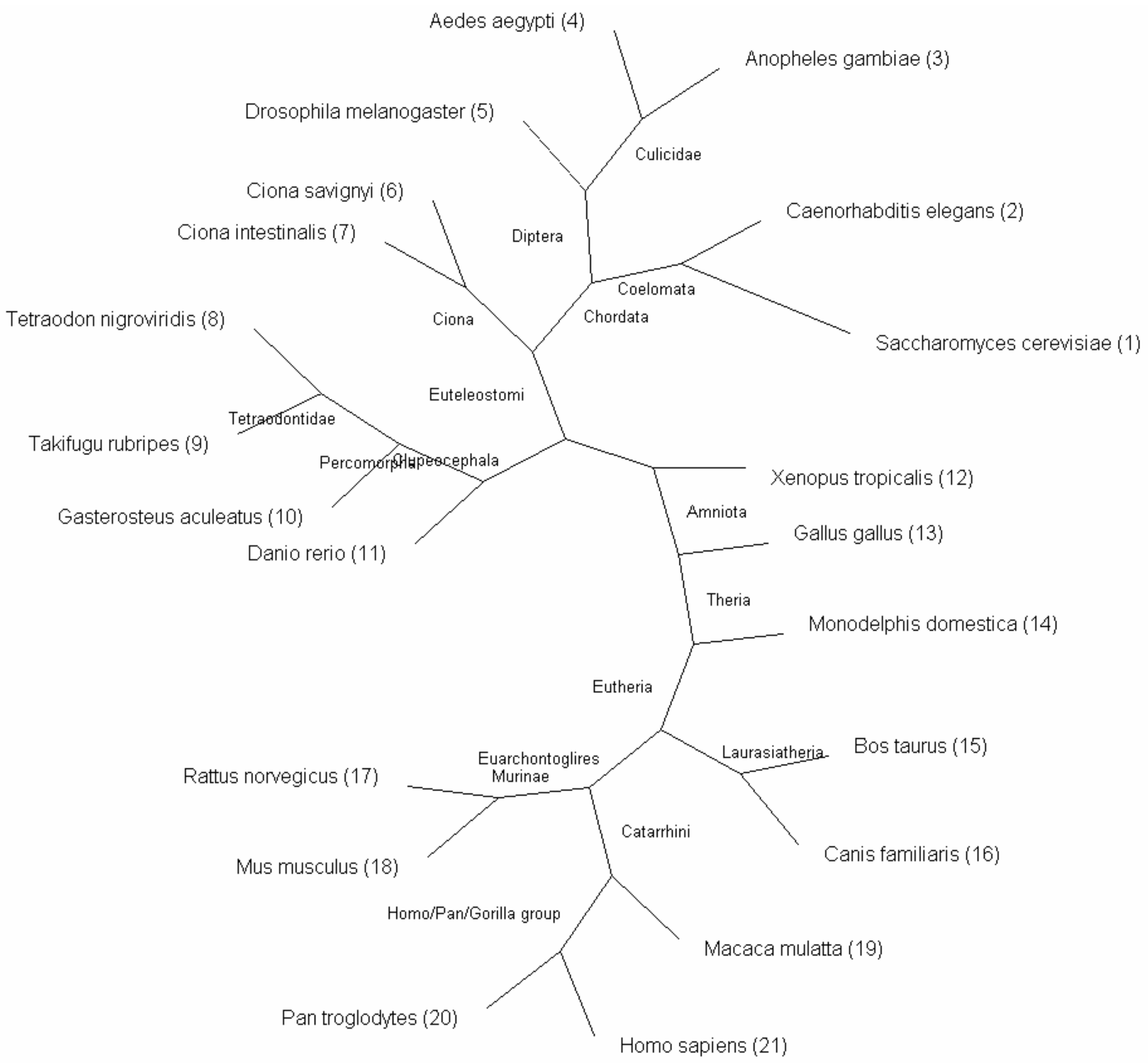
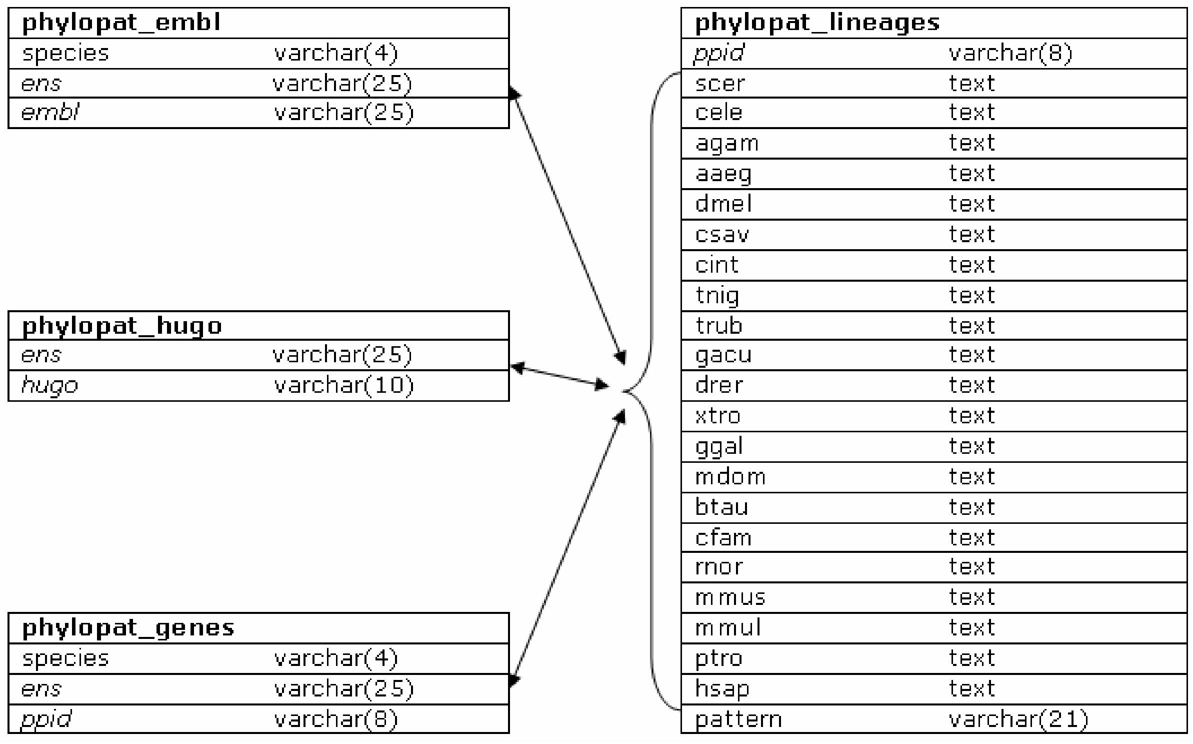


Figure 1. Phylogenetic tree of all species present in PhyloPat

This is the unrooted NCBI Taxonomy tree of all species available in Ensembl and PhyloPat. The numbers are the order in which the species are shown on the PhyloPat results pages. A phylogram version of this tree is available through the website.



phylopat_embl	
species	varchar(4)
ens	varchar(25)
embl	varchar(25)

phylopat_hugo	
ens	varchar(25)
hugo	varchar(10)

phylopat_genes	
species	varchar(4)
ens	varchar(25)
ppid	varchar(8)

phylopat_lineages	
ppid	varchar(8)
scer	text
cele	text
agam	text
aaeg	text
dmel	text
csav	text
cint	text
tnig	text
trub	text
gacu	text
drer	text
xtro	text
ggal	text
mdom	text
btau	text
cfam	text
rnor	text
mmus	text
mmul	text
ptro	text
hsap	text
pattern	varchar(21)

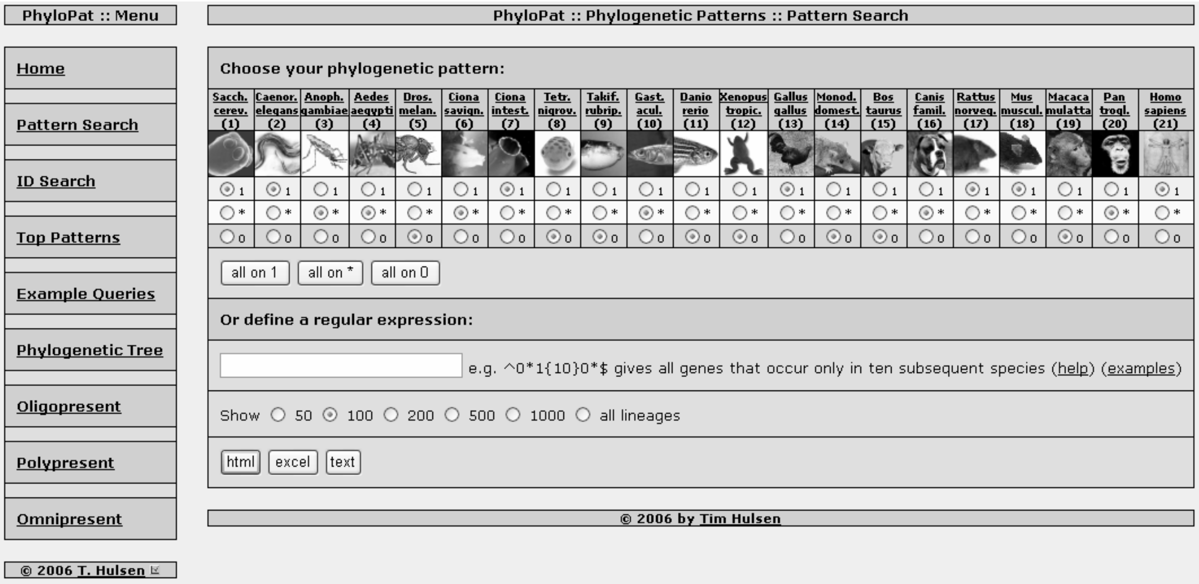
Figure 2. The PhyloPat database scheme

The database scheme shows all four tables used in the application. Table names are in bold, primary keys are in italic. Links between fields are shown with arrows. The left side of each column shows the field names, the right side shows the field types.

4.4 Utility & Discussion

4.4.1 Utility

We developed an intuitive web interface (figure 3) named PhyloPat to query a MySQL database containing these phylogenetic lineages and derived phylogenetic patterns. As input a phylogenetic pattern is used, generated by clicking a set of radio buttons or by typing a regular expression, or a list of Ensembl or EMBL identifiers. The application of MySQL regular expressions provides enhanced querying. The output can be given in HTML, Excel or plain text format. A link to the FatiGO web interface has been incorporated in the HTML output, creating easy access to functional information. Each phylogenetic lineage can be viewed separately by clicking the PhyloPat ID (PPID). This view gives all Ensembl IDs within the phylogenetic lineage plus the HUGO [14] gene names. The web interface also provides some example queries, the 100 most occurring patterns, and numerical overviews of lineages that are present in 1) all species 2) almost all species and 3) only one or two species. Finally, a phylogenetic tree of all included species is provided, through which each branch can be selected to view a list of branch-specific genes. This tree can be downloaded in PHYLIP [15] format.



PhyloPat :: Phylogenetic Patterns :: Pattern Search

Choose your phylogenetic pattern:

Sacch. cerev. (1)	Caenor. elegans (2)	Anoph. gambiae (3)	Aedes aegypti (4)	Dros. melan. (5)	Ciona savign. (6)	Ciona intest. (7)	Tetr. nigrov. (8)	Takif. rubrip. (9)	Gast. acul. (10)	Danio rerio (11)	Xenopus tropic. (12)	Gallus gallus (13)	Monod. domest. (14)	Bos taurus (15)	Canis famil. (16)	Rattus norveg. (17)	Mus muscul. (18)	Macaca mulatta (19)	Pan trogl. (20)	Homo sapiens (21)
● 1	● 1	○ 1	○ 1	○ 1	○ 1	● 1	○ 1	○ 1	○ 1	○ 1	○ 1	● 1	○ 1	○ 1	○ 1	● 1	● 1	○ 1	○ 1	● 1
○ *	○ *	● *	● *	○ *	● *	○ *	○ *	○ *	● *	○ *	○ *	○ *	○ *	○ *	● *	○ *	○ *	○ *	● *	○ *
○ 0	○ 0	○ 0	○ 0	● 0	○ 0	○ 0	● 0	● 0	○ 0	● 0	● 0	○ 0	● 0	● 0	○ 0	○ 0	○ 0	● 0	○ 0	○ 0

Or define a regular expression:

e.g. ^0*1{10}0*$ gives all genes that occur only in ten subsequent species (help) (examples)

Show ○ 50 ● 100 ○ 200 ○ 500 ○ 1000 ○ all lineages

Figure 3. The PhyloPat web interface (Pattern Search tab)

The web interface has the menu on the left and the input/results page on the right. On the pattern search page, the user can generate a phylogenetic pattern by clicking a radio button for each species. 1 = present, * = present/absent, 0 = absent. The buttons directly below put all 21 species on the corresponding mode. MySQL regular expressions offer the possibility of advanced querying. The user can choose to show any number of lineages and choose the output format: HTML, Excel or plain text. *Color version on page 158.*

4.4.2 Omnipresent genes

An analysis of all lineages with the phylogenetic pattern '111111111111111111111' (or MySQL regular expression '^1+$') gives a list of 'omnipresent' genes, i.e. present in all 21 species. We found 1001 omnipresent genes, which are most likely involved in important functions, since they are present in all species. Figure 4 shows the GO annotation [16] for all 2185 human genes within these omnipresent phylogenetic lineages, generated by FatiGO [17]. When human genes are present in the output, FatiGO can be queried by clicking a button below the output. To compare the results, we also show the GO annotation for the complete set of human genes (31,718 in Ensembl v40). Lines are drawn between similar GO classifications, to facilitate easy comparison between the omnipresent genes and all human genes. It is clear from the 6th level GO biological process annotation (figure 4a) that omnipresent genes are less often involved in transcription compared to a human gene chosen at random, but more often in cellular protein metabolism and establishment of cellular localization. We suggest that the process of transcription does not need that many genes in the 'lower' species, but in the 'higher' species, like human, many transcription related gene families have expanded ([18], table 1). Analysis of the 6th level GO molecular functions (figure 4b) shows that many omnipresent genes have ATP binding or pyrophosphatase activity, while the human gene set consists for almost 10% of genes with rhodopsin-like receptor activity. The latter is due to the fact that the GPCR class A family has expanded greatly in mammals ([19], table 2). Finally, the 6th level GO cellular components (figure 4c) show that a lesser fraction of the omnipresent genes are integral to the plasma membrane.

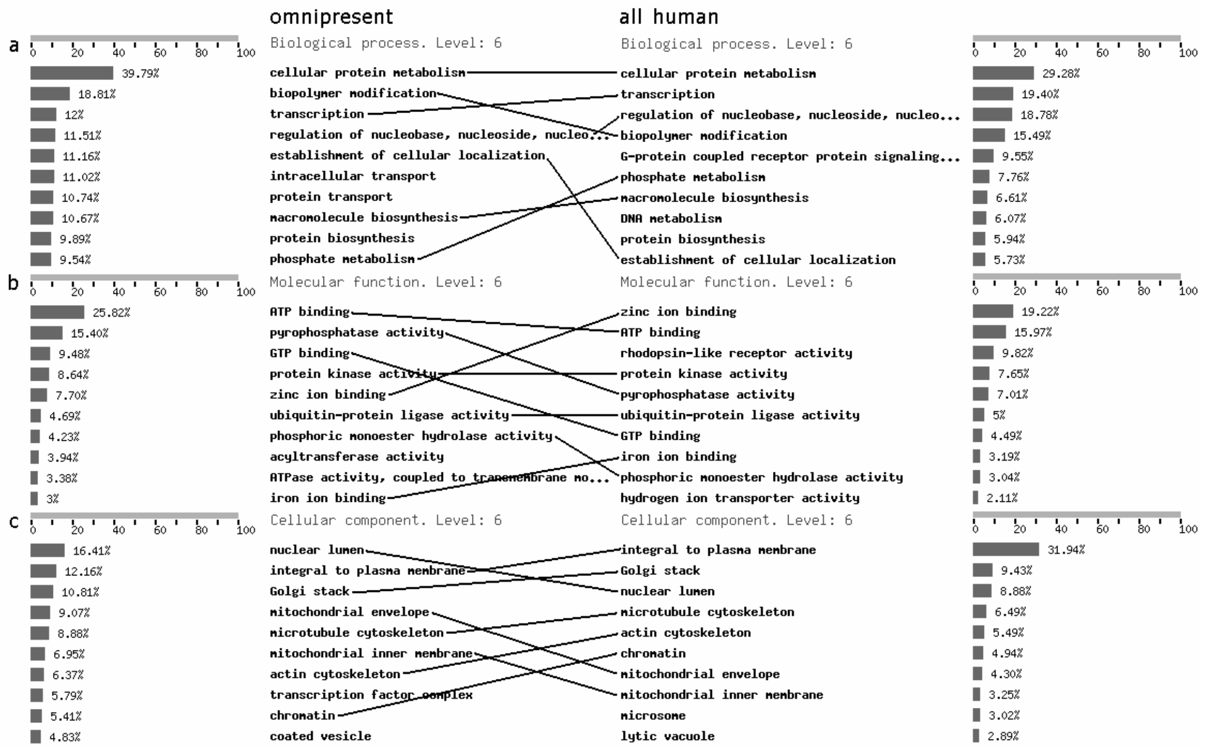


Figure 4. Gene Ontology annotations of 1) omnipresent and 2) all human genes

The left side shows the Gene Ontology annotations for all 2,185 human genes in omnipresent lineages. The right side shows the Gene Ontology annotations for all 31,718 human genes, used as a reference set. Lines are placed between equal annotations for easy comparisons between the left and the right side. **(a)** 6th level GO Biological Processes. **(b)** 6th level GO Molecular Functions. **(c)** 6th level GO Cellular Components.

4.4.3 Oligopresent genes

The distribution of 'oligopresent' genes (genes that exist in only one/two species) can be used to determine which species are evolutionary most related, as the number of shared genes, that are absent in other species, can be used as a measure for the phylogenetic distance [20]. It is apparent that are the closest relatives are *C. savignyi* and *C. intestinalis* (1737 oligopresent genes), followed by *T. nigroviridis* and *T. rubripes* (1572 oligopresent genes) and *A. gambiae* and *A. aegypti* (1058 oligopresent genes). These results correspond perfectly with the current opinion on evolutionary relationships. It should also be noted that the number of genes present in only one species is this high because of the incomplete orthology information contained in the EnsMart database. This will improve with each new Ensembl release, as orthology information and functional annotation are expanded in each release.

4.4.4 Polypresent genes

A second measure for evolutionary relatedness is the distribution of 'polypresent' genes: genes that are missing in only one or two species. *S. cerevisiae* has the highest number of missing polypresent genes: 961 polypresent genes do not occur in *S. cerevisiae* only, and 854 polypresent genes are not present in *S. cerevisiae* and a second species. Other high-scoring pairs include both *Ciona* species (47 absent polypresent genes) and the combination of one of these *Ciona* species with *G. gallus* (16 and 14 absent polypresent genes). The relatively high number for the latter pair is striking, because these species are not closely related. One would suspect such a high number only for two species that are relatively closely related, which is the case for the two *Ciona* species.

4.4.5 Case study: Hox genes

As a case study we used the highly researched and from an evolutionary point-of-view very interesting Hox genes. First, we searched the Ensembl database for human genes with the term 'hox' in the annotation. We found 44 genes, which were entered into PhyloPat. The output is shown in Table 1. The lists of Ensembl IDs have been replaced by the number of IDs. 32 phylogenetic lineages were found, one of which were already present in *C. elegans*: PP022041. This lineage contains the Msh homeobox-like proteins. PP024984 and PP027791, containing the HOXC4 and TLX lineages, are only found in the Coelomata: *A. gambiae* and further. No less than 22 lineages originated in the early vertebrates, presented by *T. nigroviridis*. HOXD12 and HOXB13 are only present in mammals.

Table 1. Phylogenetic lineages containing human HOX cluster genes

ppid	sc	ce	ag	aa	dm	cs	ci	tn	tr	ga	dr	xt	gg	md	bt	cf	rn	mm	mm	pt	hs	pattern	gene
PP022041	0	1	1	1	1	1	1	3	6	5	6	2	2	3	3	2	3	3	2	2	2	011111111111111111111	MSX1,2
PP024984	0	0	1	0	0	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	001000011111001111111	C4
PP027791	0	0	1	1	1	0	0	2	3	3	4	3	2	3	3	3	3	3	3	3	3	001110011111111111111	TLX1,2,3
PP049478	0	0	0	0	0	0	2	2	1	1	5	3	1	1	2	3	2	2	2	2	3	000000111111111111111	B8,C8,D8
PP053824	0	0	0	0	0	0	0	1	1	1	2	0	0	1	0	1	0	1	0	1	1	000000011110010101011	D11
PP053827	0	0	0	0	0	0	0	2	2	2	1	1	1	1	1	1	1	1	1	1	1	000000011111111111111	A10
PP053828	0	0	0	0	0	0	0	2	1	1	1	1	2	1	2	1	2	2	2	2	2	000000011111111111111	C13,D13
PP053829	0	0	0	0	0	0	0	6	3	3	4	1	1	2	2	2	2	2	2	2	2	000000011111111111111	A1,B1
PP053830	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	1	1	1	1	1	1	000000011110010111111	B4
PP053832	0	0	0	0	0	0	0	2	1	1	1	1	0	1	1	1	1	1	1	1	1	000000011111011111111	A5
PP053833	0	0	0	0	0	0	0	2	1	1	1	0	1	1	1	1	1	1	0	1	1	000000011110111111011	B2
PP053834	0	0	0	0	0	0	0	3	1	1	0	1	0	1	1	1	1	1	1	1	1	000000011101011111111	D3
PP053835	0	0	0	0	0	0	0	2	1	1	1	0	1	1	1	1	1	1	1	0	1	000000011110111111101	A9
PP053836	0	0	0	0	0	0	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	000000011111111111111	A3
PP053838	0	0	0	0	0	0	0	2	1	1	1	0	1	0	1	1	1	1	1	1	1	000000011110101111111	C12
PP053839	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	000000011111111110111	D4
PP053840	0	0	0	0	0	0	0	2	1	1	1	1	2	0	1	0	1	1	1	0	1	000000011111101011101	C11
PP053842	0	0	0	0	0	0	0	4	3	2	2	1	1	1	1	1	1	1	1	1	1	000000011111111111111	A13
PP053844	0	0	0	0	0	0	0	3	2	2	3	1	0	1	1	1	1	1	1	1	1	000000011111011111111	B5
PP053845	0	0	0	0	0	0	0	2	1	1	1	1	1	1	1	1	1	1	0	1	1	000000011111111111011	B3
PP053846	0	0	0	0	0	0	0	2	1	1	2	1	1	1	1	1	1	1	1	1	1	000000011111111111111	D10
PP053847	0	0	0	0	0	0	0	2	2	2	1	1	1	1	1	1	1	1	1	1	1	000000011111111111111	A2
PP053849	0	0	0	0	0	0	0	3	4	1	5	1	1	3	2	3	3	3	3	2	3	000000011111111111111	A6,B6,C6
PP053853	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	000000011101111111011	A4
PP053854	0	0	0	0	0	0	0	3	2	2	5	2	2	2	3	1	3	3	2	1	3	000000011111111111111	B9,C9,D9
PP053858	0	0	0	0	0	0	0	1	1	1	2	0	0	1	1	1	1	1	1	1	1	000000011110011111111	A11
PP070659	0	0	0	0	0	0	0	0	0	1	2	1	2	1	2	2	2	2	2	2	2	000000000111111111111	A7,B7
PP075622	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1	000000000010001111111	C5
PP084287	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	000000000001101111111	C10
PP085049	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	1	1	1	000000000001011011111	D1
PP087941	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	000000000000111011111	D12
PP089685	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	000000000000111111111	B13

Striking observations can be made with the fish species: all three species have significantly more Hox genes than the mammals. *T. nigroviridis*, for example, has 57 genes in this lineage, while *M. domestica* has only 35. These numbers correspond well with the fact that Teleost fish have at least seven Hox clusters, whereas mammals have only four [21]. Mammals also have less Hox genes per cluster, demonstrating that there has been gene loss within the Hox clusters since the evolution from a vertebrate ancestor to present-day mammals [22]. Table 2 shows the further analysis of the Hox genes using the PhyloPat output. *H. sapiens* misses the genes HOXA8, HOXB10, HOXB11, HOXC1, HOXC2, HOXC3, HOXC7, HOXD2, HOXD5, HOXD6 and HOXD7. The absence of these 11 genes is in agreement with current knowledge of human Hox genes (figure 3a of [22]). Two exceptions exist: HOXC8 instead of HOXC7, and the absence of HOXA12. The HOXA12 gene cannot be found in the other mammals either.

Table 2. Analysis of phylogenetic lineages containing human HOX cluster genes

ppid(s)	name	cluster A	cluster B	cluster C	cluster D	first sp.	position
PP053829,PP085049	HOX1	HOXA1	HOXB1		HOXD1	T. nigrov.	anterior
PP053847,PP053833	HOX2	HOXA2	HOXB2			T. nigrov.	anterior
PP053836,PP053845,PP053834	HOX3	HOXA3	HOXB3		HOXD3	T. nigrov.	PG3
PP053853,PP053830,PP024984,PP053839	HOX4	HOXA4	HOXB4	HOXC4	HOXD4	A. gamb.	central
PP053832,PP053844,PP075622	HOX5	HOXA5	HOXB5	HOXC5		T. nigrov.	central
PP053849	HOX6	HOXA6	HOXB6	HOXC6		T. nigrov.	central
PP070659	HOX7	HOXA7	HOXB7			G. acul.	central
PP049478	HOX8		HOXB8	HOXC8	HOXD8	C. intest.	central
PP053835,PP053854	HOX9	HOXA9	HOXB9	HOXC9	HOXD9	T. nigrov.	posterior
PP053827,PP084287,PP053846	HOX10	HOXA10		HOXC10	HOXD10	T. nigrov.	posterior
PP053858,PP053840,PP053824	HOX11	HOXA11		HOXC11	HOXD11	T. nigrov.	posterior
PP053838,PP087941	HOX12			HOXC12	HOXD12	T. nigrov.	posterior
PP053842,PP089685,PP053828	HOX13	HOXA13	HOXB13	HOXC13	HOXD13	T. nigrov.	posterior
PP027791	TLX	TLX1	TLX2	TLX3		A. gamb.	
PP022041	MSX	MSX1	MSX2			C. eleg.	

4.4.6 Functional annotation

PhyloPat can be used for annotation of genes with unknown functions. When a gene with unknown function is clustered in a certain phylogenetic lineage, the function of other genes in that lineage can be assigned to the gene with unknown function. For example, the PP001723 lineage [23] contains a number of genes that have an unknown function, under which the ENSANGG00000008970 gene from *A. gambiae* and the ENSCING00000000880 gene from *C. intestinalis*. By using the orthology information provided by Ensembl and the PhyloPat clustering into one lineage, we can see that all of these genes are connected to the human gene KLHDC4. This function can now be assigned to the genes with unknown function.

4.4.7 Discussion

The above examples show that PhyloPat is useful in evolutionary studies and gene annotation. It continues on the concept of phylogenetic pattern tools like EPPS [2], and on gene databases like TreeFam [6] and Homogen [5]. The originality of PhyloPat lies in the combination of these two aspects: phylogenetic pattern querying and gene family databases. In PhyloPat it is possible to determine a species set that should be included (1), a species set that should be excluded (0) and a species set which presence is indifferent (*). This, and the use of regular expression queries, enables quite complicate phylogenetic patterns searches and clustering. For example, with PhyloPat it is quite easy to find two sets of genes that have completely anti-correlating patterns (like '001111100011000000000' and '110000011100111111111'). Some of these genes from the different sets might turn out to be analogous, i.e. performing the same function but having different ancestor genes. Such kind of analysis is very hard to do with TreeFam or Hogenom. Furthermore, we aim to provide an easy-to-use web interface in which the Ensembl database can be queried using phylogenetic patterns. In just one second, users can see which gene families are present in a certain species set but missing in another species set. The output of our application can be easily analyzed by the FatiGO tool, like we demonstrated in figure 4. Finally, PhyloPat has the advantage of only relying on the Ensembl database. Treefam and Hogenom use a wide range of gene and protein databases, each with their own standards and methodologies. By using only the Ensembl database (considered by many to be the standard genome database) as input, we create a non-redundant database, through which it is possible to easily study lineage-specific expansions of gene families.

4.5 Conclusion

The analyses of the oligopresent, polypresent and omnipresent genes, as well as the small case study of the Hox genes, are just a few examples of what can be done with phylogenetic patterns in general and PhyloPat in particular. Using this tool, it is easy to find genes that e.g. occur for the first time in vertebrates, occur only in a specific number of species, or are unique for a certain species. It will be of help in the annotation of genes with unknown functions. By comparing the genes in lineages with anticorrelating patterns, it will also help finding analogous genes. PhyloPat will be completely recalculated with each major Ensembl release to ensure up-to-date and reliable phylogenetic lineages.

4.6 Availability and requirements

PhyloPat is freely available at http://www.cmbi.ru.nl/phylopat/.

4.7 Acknowledgements

This work was supported financially by NV Organon. The authors like to thank Erik Franck, Nanning de Jong and Arnold Kuzniar for their suggestions. We thank Greer Wilson for critically reading this manuscript.

4.8 References

1. Natale DA, Galperin MY, Tatusov RL, Koonin EV: Using the COG database to improve gene recognition in complete genomes. *Genetica* 2000, 108(1):9-17.
2. Reichard K, Kaufmann M: EPPS: mining the COG database by an extended phylogenetic patterns search. *Bioinformatics* 2003, 19(6):784-785.
3. Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS: OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 2006, 34(Database issue):D363-368.
4. Dehal PS, Boore JL: A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* 2006, 7:201.
5. Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G: Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 2005, 21(11):2596-2603.
6. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L et al: TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* 2006, 34(Database issue):D572-580.
7. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T et al: Ensembl 2006. *Nucleic Acids Res* 2006, 34(Database issue):D556-561.
8. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: EnsMart: a generic system for fast and flexible access to biological data. *Genome Res* 2004, 14(1):160-169.
9. Ensembl orthology and paralogy prediction pipeline [http://www.ensembl.org/info/data/compara/homology_method.html]
10. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006, 7(4):R31.
11. Edgar RC: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004, 5:113.
12. Guindon S, Gascuel O: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003, 52(5):696-704.
13. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2000, 28(1):10-14.
14. Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ: The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res* 2006, 34(Database issue):D319-321.
15. PHYLIP (Phylogeny Inference Package) [http://evolution.genetics.washington.edu/phylip.html]
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25(1):25-29.
17. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 2004, 20(4):578-580.

18. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV et al: TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003, 31(1):374-378.
19. Fredriksson R, Schioth HB: The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol* 2005, 67(5):1414-1425.
20. Korbel JO, Snel B, Huynen MA, Bork P: SHOT: a web server for the construction of genome phylogenies. *Trends Genet* 2002, 18(3):158-162.
21. Wagner GP, Amemiya C, Ruddle F: Hox cluster duplications and the opportunity for evolutionary novelties. *Proc Natl Acad Sci U S A* 2003, 100(25):14603-14606.
22. Minguillon C, Gardenyes J, Serra E, Castro LF, Hill-Force A, Holland PW, Amemiya CT, Garcia-Fernandez J: No more than 14: the end of the amphioxus Hox cluster. *Int J Biol Sci* 2005, 1(1):19-23.
23. PhyloPat lineage PP001723 [http://www.cmbi.ru.nl/pw/phylopat/40/phylopat.php?ppid=PP001723]

Chapter 5

Evolution of the immune system from model organism to man

Tim Hulsen, Wilco W.M. Fleuren, Hindrik H.D. Kerstens, Martien A.M. Groenen and Peter M.A. Groenen

Manuscript in preparation

5.1 Abstract

The immune system is of major importance since it protects metazoans from infection by pathogenic organisms. Throughout evolution, two major branches have originated: innate and adaptive immunity. Innate immunity uses the genetic memory of germline-encoded receptors to recognize the molecular patterns of common pathogens. Adaptive immunity is a complex system by which the body learns to recognize a pathogen's unique antigens and builds an antigen specific response to destroy it. The innate immune system exists in a wide range of metazoans, whereas the adaptive immune system is only present in jawed vertebrates. Both the innate and the adaptive immune system are intensively studied by scientists working in the field of drug discovery, since numerous drugs are active in immunologic pathways. However, immunologic drug discovery is difficult since there are sometimes large differences in drug response between model organisms and man. These differences might be explained by studying the evolution of genes involved in the immune system.

In this article we present an overview of the evolution of the immune system from several model organisms to man, using whole-genome data from a wide range of species. First, we use the Ensembl database and the PhyloPat application to create phylogenetic lineages related to the immune system. These lineages are available through our web application ImmunoPhyle at http://www.cmbi.ru.nl/immunophyle. Second, we identify lineage-specific expansions and deletions within the vertebrate immune system. This identification is made easier because of the use of genome data instead of proteome data: our view is not disturbed by alternative transcripts or isoforms. Third, we show the evolutionary differences between the innate and the adaptive immune system. Finally, we zoom in on several interesting families and show how our data can be mapped onto pathways. We conclude that our analyses can be used to explain differences in (immunologic) drug responses between model organisms and man, but that in most cases a combination of orthology data, expression data, protein interaction data and structural data is needed to find a sufficient explanation.

5.2 Introduction

5.2.1 Immunity: innate and adaptive

The immune system protects metazoans from infection by pathogenic organisms. It is divided into two major branches, termed innate and adaptive immunity [1]. Innate immunity uses the genetic memory of germline-encoded receptors to recognize the molecular patterns of common pathogens. Adaptive immunity, closely related to somatic memory, is a complex system by which the body learns to recognize a pathogen's unique antigens and builds an antigen specific response to destroy it. The effective development of the overall immune response depends on careful interplay and regulation between innate and adaptive immunity. The innate immune system exists in a wide range of metazoans [2], whereas the adaptive immune system is only present in jawed vertebrates or 'gnathostomes' [3], although recent findings suggest that this line might not be as solid as thought before [4, 5].

5.2.2 Innate immunity

The innate immune response fights infections from the moment of first contact and is the fundamental defensive weapon of multicellular organisms [6]. The family of Toll receptors has a crucial role in immune defence. Studies in fruitflies and in mammals reveal that the defensive strategies of invertebrates and vertebrates are highly conserved at the molecular level, which raises the exciting prospects of an increased understanding of innate immunity. However, the function of Toll receptors differs between mammals and *Drosophila*, raising intriguing questions about the mechanisms of Toll signal reception and the relationship between inflammation and development. Many human diseases result from the failure of processes in the innate immunity response, either caused by a primary defect or by medical treatment. The innate system can be subdivided into the afferent (or sensing) arm and the efferent (or effector) arm, each of which can be further divided into cellular and humoral components [7].

5.2.3 Adaptive immunity

The adaptive immune system originated approximately 500 million years ago, in vertebrate species [8]. In fact, there were two different adaptive immune systems that evolved convergently: one in jawed vertebrates, and one in jawless fish. While the superfamily of jawed vertebrates expanded greatly in the past 500 million years, the only jawless fish that exist nowadays are the lampreys and the hagfish [9]. Therefore, when referring to ‘the adaptive immune system’, usually the adaptive immunity evolved in jawed vertebrates is meant. This adaptive immune system uses somatically rearranged antigen receptor genes to create receptors for virtually any antigen [10]. The adaptive immune response is slower but more flexible than the innate immune response, and is able to combat infections that have evolved to evade innate responses. The adaptive immune system has the capacity to recognize and respond to virtually any protein or carbohydrate imaginable.

5.2.4 Immunogenomics

Several attempts have been made to identify immune-related genes in single species, such as chicken [11], mouse [12] and man [13]. These studies are very useful in providing answers for species-specific questions concerning the immune system. For example, thanks to immunogenetic studies following the completion of the Human Genome Project, we now know that the human immune sub-genome consists of around 1562 genes (about 7% of the total human genome) [13], not including the immunoglobulin (Ig) superfamily which makes up over 2% of human genes, possibly constituting the largest gene family in the human genome [14]. However, these studies do not tell anything about the origin and evolution of this superfamily. With more and more whole genome data becoming available, from both vertebrates and invertebrates, it is now possible to study the evolution of the immune system through different species. Genome-wide approaches to study the immune system are known as immunogenomics [15] or immunomics [16]. In this article we present an overview of the evolution of the immune system in the vertebrate lineage, using whole-genome data from a wide range of species. First, we identify lineage-specific expansions and deletions within the vertebrate immune system. This identification is made easier because of the use of genome data instead of proteome data: our view is not disturbed by alternative transcripts or isoforms. Second, we show the evolutionary differences between the innate and the adaptive immune system. Finally, we zoom in on several interesting families and show how our data can be mapped onto pathways.

5.3 Methods

5.3.1 Phylogenetic lineages

We used the Ensembl database [17], version 41, as a starting point for our immunogenomics analysis. This database contains in total 553,721 genes from 26 species: 1 yeast, 6 invertebrate animals, 7 vertebrate non-mammals and 12 mammals, under which numerous species often used as model organisms for man: *D. melanogaster, M. musculus, R. norvegicus* and *M. mulatta*. A phylogenetic tree of these species can be viewed in figure 1. We built phylogenetic lineages, i.e. orthologous groups, using a simple single linkage clustering, in the same way as for the web application PhyloPat [18]. In order to get a immune-specific data set, we gathered all HUGO [19] gene names included in the Immunogenetic Related Information Source (IRIS, [13]). All phylogenetic lineages connected to one or more of the 1551 immunologic HUGO names were stored in a separate database, named ImmunoPhyle. This database now includes 18,933 genes from the 26 species, including 1,157 genes from *H. sapiens*. Results are displayed in order from the 'lowest' species S. cerevisiae to the 'highest' species *H. sapiens* ('low'/'high' corresponding to the longest/shortest evolutionary distance to man).

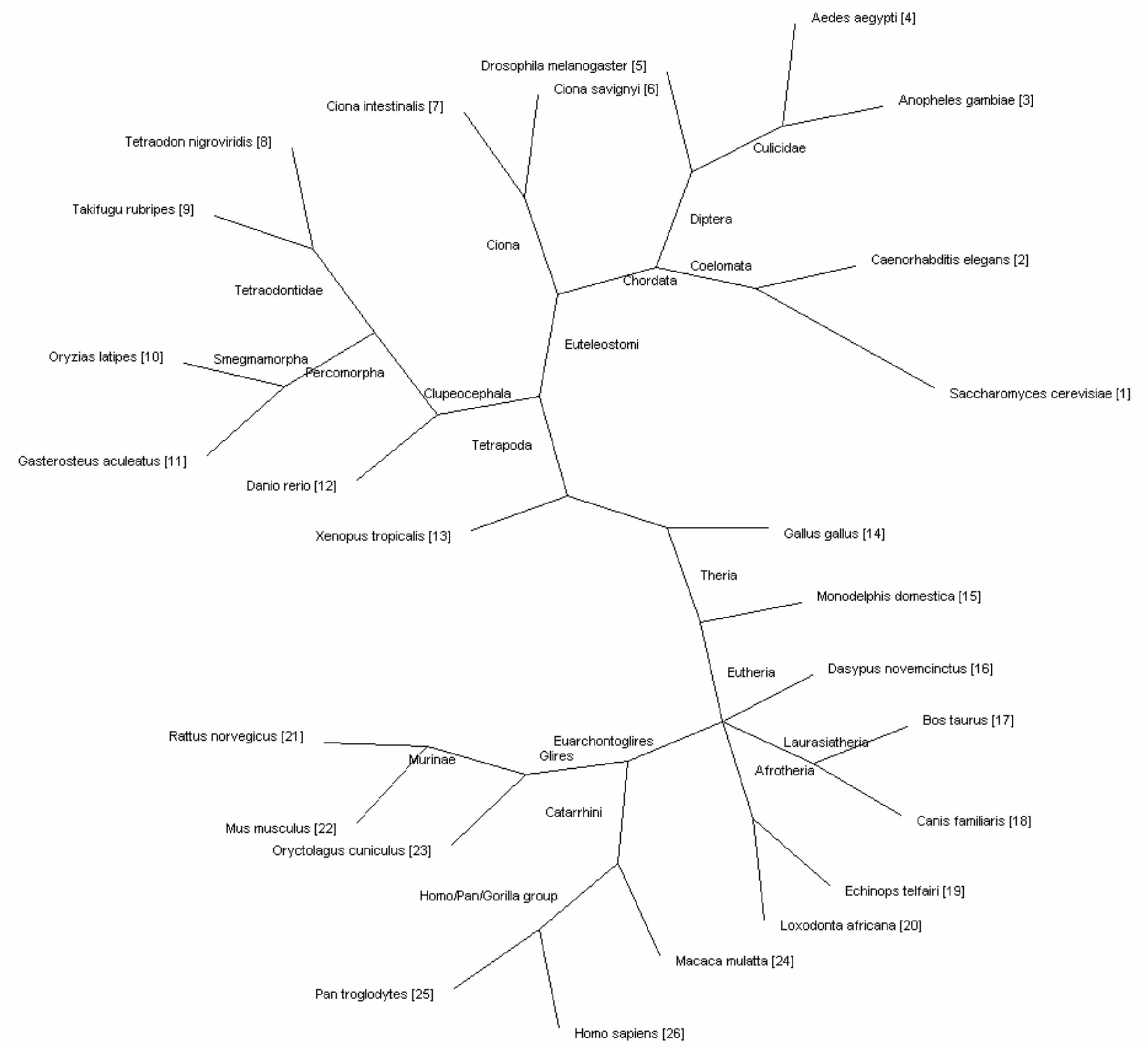


Figure 1. Phylogenetic tree of the 26 species included in our analysis
Unrooted phylogenetic tree of the 26 species included in our analysis, created by the NCBI Taxonomy database [20] and TreeView [21].

5.3.2 IRIS classification

We make use of the classification into 22 categories provided by the IRIS database. Table 1 shows these categories, together with the number of HUGO gene names, the number of orthologous groups and the number of genes within these categories. All information is available through our ImmunoPhyle web application (http://www.cmbi.ru.nl/immunophyle/). Please note that each HUGO gene name can be linked to multiple IRIS categories.

Table 1. The IRIS categories linked to the phylogenetic lineages

Nr.	Abbrev.	Description	# HUGO IDs	# ImmunoPhyle lineages	# Genes
1	InImm	Innate Immunity	638	272	8640
2	Inflm	Inflammation	314	117	4568
3	Chmtx	Chemotaxis	192	54	2374
4	Phago	Phagocytosis	37	17	890
5	Compl	Complement	62	33	958

6	Cy_Ch	Cytokines and Chemokines	261	109	2947
7	AdImm	Adaptive Immunity	422	140	4983
8	ClRsp	Cellular Response	145	63	2358
9	HmRsp	Humoral Response	98	34	1087
10	BMImm	Barrier and Mucosal Immunity	45	18	713
11	Devlp	Development of Immune System	130	50	2044
12	AgPrc	Antigen Processing	148	31	830
13	PtSig	Immune Pathway or Signalling	470	224	8245
15	Recpt	Receptor	246	118	3506
16	IndIm	Induced by Immunomodulator	200	86	3487
20	ImDef	Involved in Immunodeficiency	71	30	1013
21	AutIm	Involved in Autoimmunity	44	19	530
22	ExpIT	Expressed Primarily in Immune Tissues	332	134	3970
23	Other	Other	107	43	1843
25	InKil	Innate NK Killing	82	33	1015
26	RlDis	Related to Disease	172	91	3141
27	Coagl	Coagulation	111	51	2624
0	All	All immunologic lineages	1542	585	18933

5.4 Results

5.4.1 Expansions and deletions

Table 2 shows how many genes are linked to each category, for each of the 26 species in our dataset. From this table, it is obvious that the immune system is largely restricted to vertebrates: *T. nigroviridis*, the first vertebrate in the list, contains almost four times as many immunorelated genes as *C. intestinalis*, the last non-vertebrate in the list. This can also be concluded from figure 2, which shows an analysis of the species occurrence in the phylogenetic lineages. The largest differences can be seen in the transition from invertebrates (*C. intestinalis*) to vertebrates (*T. nigroviridis*) and from non-mammals (*G. gallus*) to mammals (*M. domesticus*). Moreover, this figure shows that *D. novemcinctus*, *L. africana* and *O. cuniculus* have a large number of deletions. This probably points to the lesser quality of the genome assembly rather than to any real evolutionary deletions. Expansions can also be viewed easily using ImmunoPhyle. The largest expansions per species can be seen in table 3.

Table 2. Numbers of genes per category and per species

Cat.	**Sc**	**Ce**	**Ag**	**Aa**	**Dm**	**Cs**	**Ci**	**Tn**	**Tr**	**Ol**	**Ga**	**Dr**	**Xt**	**Gg**	**Md**	**Dn**	**Bt**	**Cf**	**Et**	**La**	**Rn**	**Mm**	**Oc**	**Mm**	**Pt**	**Hs**	**Total**
InImm	17	51	77	89	81	81	93	351	355	339	351	420	304	295	466	355	566	435	416	384	517	571	539	384	535	568	8640
Inflm	13	34	45	57	43	53	55	202	200	194	197	237	179	150	227	200	267	221	215	197	263	302	271	194	265	287	4568
Chmtx	4	12	18	24	18	22	28	107	118	112	122	125	96	69	157	90	135	121	112	103	124	147	132	86	141	151	2374
Phago	1	4	9	10	10	8	10	46	43	41	47	50	32	31	42	34	51	45	46	42	49	58	44	33	51	53	890
Compl	0	3	13	7	7	11	19	45	41	43	43	54	37	31	50	36	58	45	48	34	60	62	55	43	54	59	958
Cy_Ch	2	11	14	20	18	18	18	122	119	124	119	148	92	120	144	106	219	173	143	133	175	187	195	143	190	194	2947
AdImm	17	44	37	40	48	59	62	212	207	204	219	253	158	170	246	188	330	260	225	223	276	315	244	324	303	319	4983
ClRsp	6	26	20	23	22	36	41	106	101	102	100	119	78	96	116	93	138	112	105	104	124	148	148	111	137	146	2358
HmRsp	3	9	8	8	9	8	9	48	46	48	45	47	37	40	49	43	60	58	55	50	61	65	76	68	65	72	1087
BMImm	0	1	10	9	15	2	4	20	25	17	27	24	18	11	33	30	68	42	38	32	48	58	47	25	52	57	713
Devlp	5	18	23	25	23	22	29	109	90	89	96	124	64	72	106	74	109	108	103	86	114	122	116	92	108	117	2044

AgPrc	3	8	9	11	11	10	12	34	31	36	38	56	22	25	39	40	49	35	37	36	39	40	63	35	54	57	830
PtSig	13	63	70	87	81	93	102	400	381	382	390	480	301	296	446	302	454	415	371	344	459	508	489	337	480	501	8245
Recpt	2	18	16	20	18	18	24	148	151	150	158	187	125	124	165	141	205	191	170	154	227	240	226	156	231	241	3506
IndIm	7	23	28	25	40	29	32	172	163	159	175	200	129	122	171	130	224	184	154	154	198	218	197	151	193	209	3487
ImDef	4	8	15	12	9	18	28	44	44	41	38	64	35	42	48	34	61	45	45	43	54	56	58	52	56	59	1013
AutIm	0	1	12	6	1	5	3	23	24	26	23	32	18	20	25	19	29	30	28	22	29	30	31	29	31	33	530
ExpIT	11	22	36	27	31	32	42	157	160	153	173	186	137	118	249	155	238	201	179	170	257	274	260	158	261	283	3970
Other	9	28	32	31	37	23	25	99	82	80	90	105	86	78	84	78	98	82	75	74	93	92	96	73	94	99	1843
InKil	1	4	9	9	6	6	8	31	37	28	32	35	44	23	38	49	70	49	52	50	74	88	72	38	80	82	1015
RlDis	6	14	32	28	32	25	34	151	143	133	144	176	115	128	159	131	190	159	153	133	170	184	184	145	182	190	3141
Coagl	5	25	36	44	33	31	33	132	123	122	124	154	114	81	124	108	141	123	121	109	145	162	139	106	138	151	2624
All	54	156	193	211	214	219	239	876	830	824	855	1015	686	685	969	740	1121	948	870	802	1070	1163	1131	818	1087	1157	18933

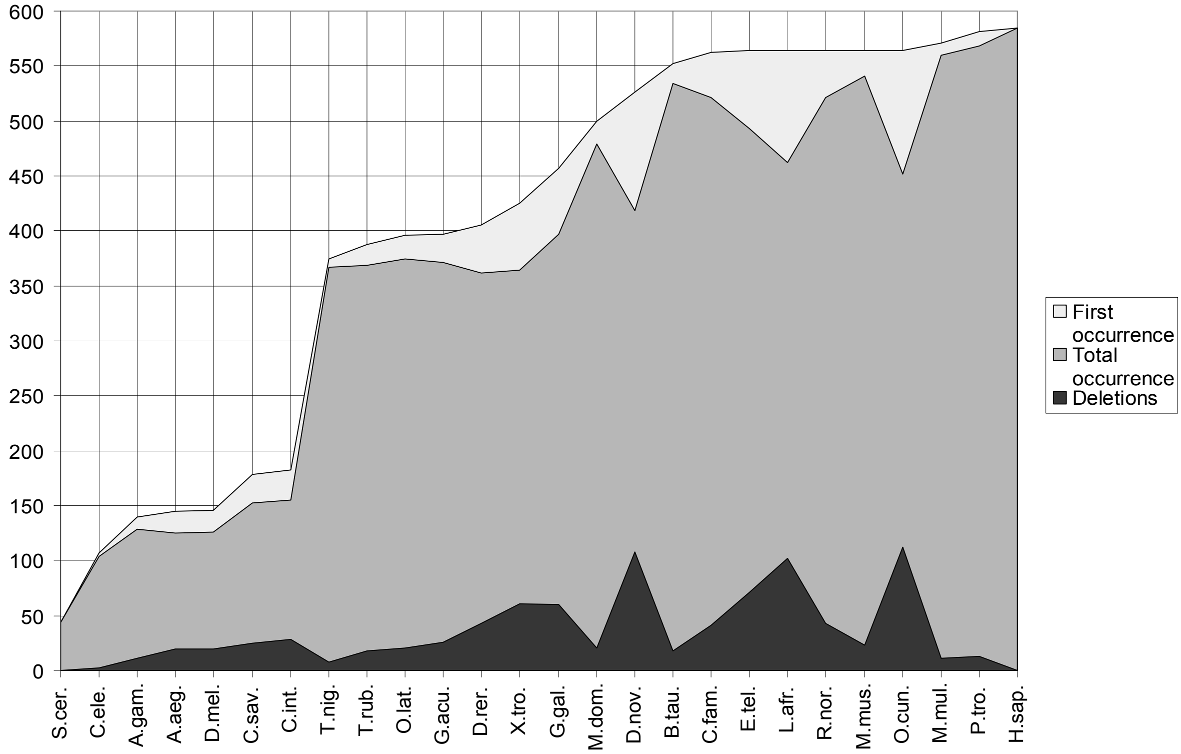


Figure 2. Analysis of species occurrence in phylogenetic lineages

Analysis of the occurrence of the 26 species in the 585 ImmunoPhyle phylogenetic lineages. Light grey: number of lineages that started in the corresponding species, or earlier. Medium grey: total number of lineages which contain one or more genes from the corresponding species. Dark grey: number of deletions in the corresponding species (dark grey = light grey – medium grey).

Table 3. Largest expansion(s) per species

Nr.	Species	#	IP	HUGO
1	S.cer.	4	IP017	HSPA1A,HSPA1B,HSPA1L,HSPA8
2	C.ele.	7	IP008	CPB2
			IP090	NR3C1
3	A.gam.	10	IP008	CPB2
4	A.aeg.	21	IP008	CPB2
5	D.mel.	14	IP008	CPB2
6	C.sav.	9	IP069	TRAF3,TRAF4,TRAF5
7	C.int.	10	IP069	TRAF3,TRAF4,TRAF5

			IP162	C6
8	T.nig.	15	IP033	MAPK8,MAPK10,MAPK11,MAPK12,MAPK13,MAPK14
9	T.rub.	12	IP035	SLC4A1
10	O.lat.	13	IP035	SLC4A1
11	G.acu.	14	IP061	ADORA1,ADORA2A,ADORA3,NCR2,PIGR,TREM1
12	D.rer.	19	IP047	A2M
13	X.tro.	16	IP229	SIGLEC5,SIGLEC6,SIGLEC7,SIGLEC8,SIGLEC9,SIGLEC10,SIGLEC11
14	G.gal.	9	IP035	SLC4A1
			IP047	A2M
15	M.dom.	54	IP463	CEACAM1,CEACAM5,CEACAM6,CEACAM8
16	D.nov.	10	IP061	ADORA1,ADORA2A,ADORA3,NCR2,PIGR,TREM1
			IP116	LYZ
			IP129	ANKRD15
			IP229	SIGLEC5,SIGLEC6,SIGLEC7,SIGLEC8,SIGLEC9,SIGLEC10,SIGLEC11
17	B.tau.	46	IP377	IFNA10,IFNA13,IFNA14,IFNA16,IFNA17,IFNA21
18	C.fam.	14	IP377	IFNA10,IFNA13,IFNA14,IFNA16,IFNA17,IFNA21
19	E.tel.	12	IP061	ADORA1,ADORA2A,ADORA3,NCR2,PIGR,TREM1
20	L.afr.	10	IP035	SLC4A1
			IP061	ADORA1,ADORA2A,ADORA3,NCR2,PIGR,TREM1
			IP147	MS4A12,MS4A4A,MS4A6A,MS4A6E,MS4A8B
21	R.nor.	17	IP530	KLRA1
22	M.mus.	23	IP061	ADORA1,ADORA2A,ADORA3,NCR2,PIGR,TREM1
23	O.cun.	18	IP291	HMGB2
24	M.mul.	16	IP377	IFNA10,IFNA13,IFNA14,IFNA16,IFNA17,IFNA21
25	P.tro.	17	IP377	IFNA10,IFNA13,IFNA14,IFNA16,IFNA17,IFNA21
26	H.sap.	16	IP061	ADORA1,ADORA2A,ADORA3,NCR2,PIGR,TREM1
			IP377	IFNA10,IFNA13,IFNA14,IFNA16,IFNA17,IFNA21
			IP463	CEACAM1,CEACAM5,CEACAM6,CEACAM8

5.4.2 Innate and adaptive immunity

The IRIS database [13] contains 22 different categories of the immune system, with each category linked to a number of HUGO [19] gene names. Figure 3 shows the overlap between the three largest categories: ‘Innate Immunity’, ‘Adaptive Immunity’ and ‘Immune Pathway or Signalling’. The overlap between all three categories (25 lineages) consists mostly of interleukins (IL), mitogen-activated protein kinases (MAPK) and killer cell immunoglobulin-like receptors (KIL). The 22 categories also differ in size: ‘Innate Immunity’ and ‘Immune Pathway or Signalling’ both contain more than 8,000 genes, while ‘Involved in Autoimmunity’ contains only 530 genes.

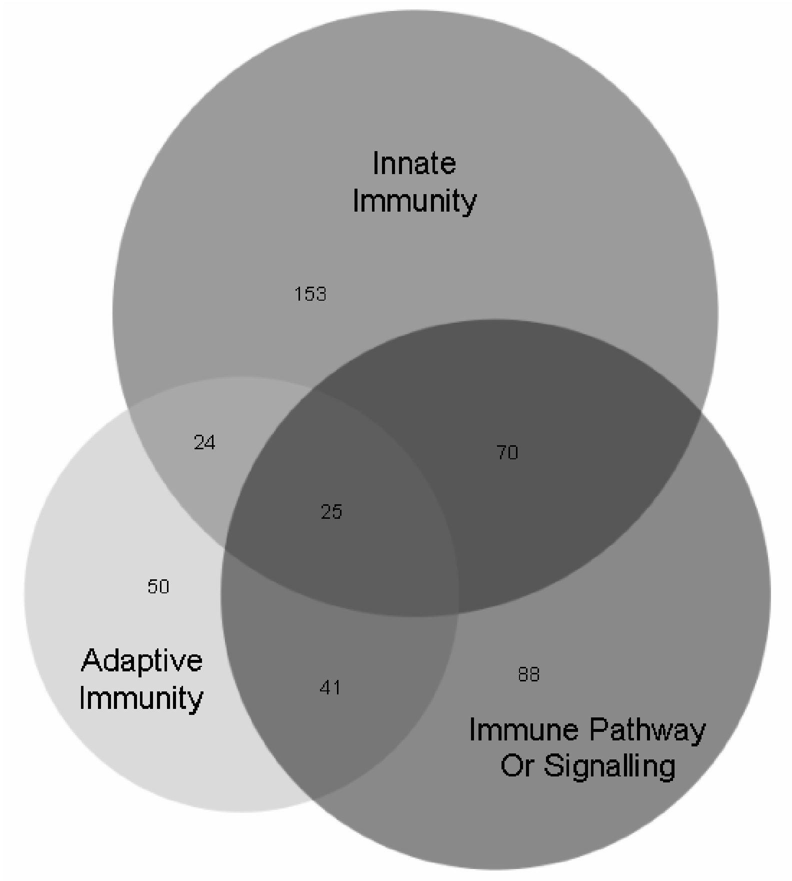


Figure 3. Venn diagram of the numbers of phylogenetic lineages linked to specific immunologic categories
Venn diagram of the numbers of phylogenetic lineages linked to 'Innate Immunity' (red), 'Adaptive Immunity' (green) and 'Immune Pathway or Signalling' (blue) and combinations of these three categories. Each surface is proportional to the number it represents, except for the overlap between all three categories. *Color version on page 158.*

As discussed in the introduction, the innate immune system should be present in all species, whereas the adaptive immune system originated in the vertebrates. Table 2 shows that the phylogenetic lineages of the adaptive immune system indeed have relatively few member genes in invertebrates: 307 genes divided over the first 7 species. The first vertebrate in the list, *T. nigroviridis*, has 212 adaptive immunity genes. However, the same trend can be seen in the innate immune system: 489 genes in the 7 invertebrate species, and 351 genes in *T. nigroviridis*. This might be caused by the anthropocentrism of our analysis: only lineages are included that are connected to a HUGO gene name, thus having a member gene in *H. sapiens*. When phylogenetic lineages are included that are not present in man, the number of innate immunity genes in invertebrates might increase drastically. In table 5, we take a closer look at the 48 adaptive immunity genes in the fruitfly *D. melanogaster*, using the annotation from the FlyBase [22] database. It shows that these genes are either encoding for proteins with unknown function that are orthologous to adaptive immunity genes in other species, or encoding for proteins that are connected to many immunologic categories, such as the heat shock proteins.

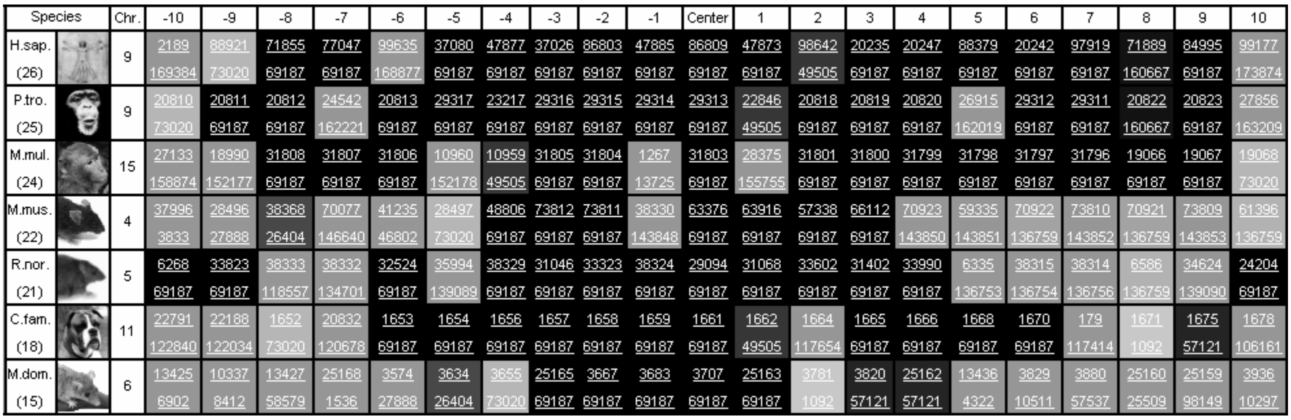
Table 5. Fruitfly genes linked to 'Adaptive Immunity' category

IPID	HUGO	Ensembl ID	Flybase ID	Flybase name
IP001	HM13	CG11840	FBgn0031260	Signal peptide protease
IP003	DPP4	CG11319	FBgn0031835	CG11319
		CG32145	FBgn0002997	omega

		CG11034	FBgn0031741	CG11034
IP007	MAEA	CG31357	FBgn0051357	CG31357
IP012	PABPC4	CG5119	FBgn0003031	polyA-binding protein
IP017	HSPA1A HSPA1B HSPA1L HSPA8	CG4264	FBgn0001219	Heat shock protein cognate 4
		CG31366	FBgn0013275	Heat-shock-protein-70Aa
		CG18743	FBgn0013276	Heat-shock-protein-70Ab
		CG31359	FBgn0013278	Heat-shock-protein-70Bb
		CG6489	FBgn0013279	Heat-shock-protein-70Bc
		CG5834	FBgn0051354	Hsp70Bbb
		CG31449	FBgn0013277	Heat-shock-protein-70Ba
		CG7756	FBgn0001217	Heat shock protein cognate 2
		CG5436	FBgn0001230	Heat shock protein 68
		CG8937	FBgn0001216	Heat shock protein cognate 1
IP020	ENTPD5	CG3059	FBgn0024947	NTPase
IP021	KPNA1	CG8548	FBgn0024889	karyopherin α1
		CG9423	FBgn0027338	karyopherin α3
IP024	MAPK1 MAPK3	CG12559	FBgn0003256	rolled
IP033	MAPK10 MAPK11 MAPK12 MAPK13 MAPK14 MAPK8	CG7393	FBgn0024846	p38b
		CG5475	FBgn0015765	Mpk2
		CG5680	FBgn0000229	basket
		CG33338	FBgn0046322	p38c
IP034	TIA1 TIAL1	CG5422	FBgn0005649	Rox8
		CG4787	FBgn0039572	CG4787
		CG12870	FBgn0039570	CG12870
IP039	PTPRCAP	CG9446	FBgn0033109	coro
IP040	CANX	CG11958	FBgn0015622	Calnexin 99A
		CG9906	FBgn0030755	CG9906
		CG1924	FBgn0030377	CG1924
IP043	LGMN	CG4406	FBgn0023545	CG4406
IP045	NFKB1 NFKB2 NFKBIA	CG5848	FBgn0000250	cactus
IP049	TNFRSF25	CG7323	FBgn0036943	CG7323
IP051	DPP8	CG3744	FBgn0039240	CG3744
IP052	MAP2K1 MAP2K2	CG15793	FBgn0010269	Downstream of raf1
IP055	TOB1	CG9214	FBgn0028397	Tob
IP061	ADORA1 ADORA2A ADORA3 NCR2 PIGR TREM1	CG9753	FBgn0039747	Adenosine receptor
IP068	ATF1 CREM	CG6103	FBgn0014467	Cyclic-AMP response element binding protein B at 17A
IP069	TRAF3 TRAF4 TRAF5	CG3048	FBgn0026319	TNF-receptor-associated factor 1
IP073	LAT	CG8428	FBgn0004571	spinster
IP085	LAMB1	CG7123	FBgn0002527	Laminin B1
IP092	CALR	CG9429	FBgn0005585	Calreticulin
IP094	ILF2	CG5641	FBgn0038046	CG5641
IP095	PACS1	CG5405	FBgn0020647	Krueppel target at 95D
IP098	MAP4K1 MAP4K2	CG7097	FBgn0034421	CG7097
IP118	PDCD4	CG10990	FBgn0030520	CG10990
IP144	MAFB	CG10034	FBgn0000964	traffic jam

5.4.3 Gene order conservation

The Ensembl database and our derivatives PhyloPat and Immunophyle offer the possibility of studying the neighbouring genes of each gene. The conservation of this gene neighbourhood, or gene order, over multiple genomes has been shown to indicate a functional association between the proteins they encode [23]. Figure 4 shows the gene neighbourhood for phylogenetic lineage IP377 or PP069187. This lineage consists of several members of the interferon alpha (IFNA) family. These IFNA genes are clustered together on almost each of the genomes, especially in *H. sapiens*, *P. troglodytes* and *M. mulatta*. This could point to very recent gene duplications, making the genes in this cluster so-called in-paralogs [24]. The genes in the direct neighbourhood of the IFNA cluster are functionally related, e.g. PP049505 (KLHL9/KLHL13), PP160667 (PTPLAD2) and PP057121 (CDKN2B/CDKN2C/CDKN2D). This kind of analysis is possible for all genes in the immune sub-genome and can, as shown here, give extra information about the evolutionary background of these genes.



Species		Chr.	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	Center	1	2	3	4	5	6	7	8	9	10
H.sap. (26)		9	2189 169384	88921 73020	71855 69187	77047 69187	99635 168877	37080 69187	47877 69187	37026 69187	86803 69187	47885 69187	86809 69187	47873 69187	98642 49505	20235 69187	20247 69187	88379 69187	20242 69187	97919 69187	71889 160667	84995 69187	99177 173874
P.tro. (25)		9	20810 73020	20811 69187	20812 69187	24542 162221	20813 69187	29317 69187	23217 69187	29316 69187	29315 69187	29314 69187	29313 69187	22846 49505	20818 69187	20819 69187	20820 69187	26915 162019	29312 69187	29311 69187	20822 160667	20823 69187	27856 163209
M.mul. (24)		15	27133 158874	18990 152177	31808 69187	31807 69187	31806 69187	10960 152178	10959 49505	31805 69187	31804 69187	1267 13725	31803 69187	28375 155755	31801 69187	31800 69187	31799 69187	31798 69187	31797 69187	31796 69187	19066 69187	19067 69187	19068 73020
M.mus. (22)		4	37996 3833	28496 27888	38368 26404	70077 146640	41235 46802	28497 73020	48806 69187	73812 69187	73811 69187	38330 143848	63376 69187	63916 69187	57338 69187	66112 69187	70923 143850	59335 143851	70922 136759	73810 143852	70921 136759	73809 143853	61396 136759
R.nor. (21)		5	6268 69187	33823 69187	38333 118557	38332 134701	32524 69187	35994 139089	38329 69187	31046 69187	33323 69187	38324 69187	29094 69187	31068 69187	33602 69187	31402 69187	33990 69187	6335 136753	38315 136754	38314 136756	6586 136759	34624 139090	24204 69187
C.fam. (18)		11	22791 122840	22188 122034	1652 73020	20832 120678	1653 69187	1654 69187	1656 69187	1657 69187	1658 69187	1659 69187	1661 69187	1662 49505	1664 117654	1665 69187	1666 69187	1668 69187	1670 69187	179 117414	1671 1092	1675 57121	1678 106161
M.dom. (15)		6	13425 6902	10337 8412	13427 58579	25168 1536	3574 27888	3634 26404	3655 73020	25165 69187	3667 69187	3683 69187	3707 69187	25163 69187	3781 1092	3820 57121	25162 57121	13436 4322	3829 10511	3880 57537	25160 25509	25159 98149	3936 10297

Figure 4. Conservation of gene order for phylogenetic lineage IP377 (IFNA)
Conservation of gene order for phylogenetic lineage IP377 or PP069187, which consists of several members of the IFNA (interferon alpha) family, for seven species: *H. sapiens*, *P. troglodytes*, *M. mulatta*, *M. musculus*, *R. norvegicus*, *C. familiaris* and *M. domestica*. For each species, the most central IFNA gene is shown next to its twenty surrounding genes on the chromosome. Black: gene belonging to the IFNA phylogenetic lineage. Color: gene belonging to phylogenetic lineage with two or more members in this figure. Grey: belonging to phylogenetic lineages with only one member in this figure ('singleton'). Only the final five/six characters of each Ensembl ID or PPID are shown. *Color version on page 159.*

5.4.4 Interleukins

The family of interleukins is a good example of how expansions and deletions within the immune system can be studied using our system of phylogenetic lineages (table 5). According to our analysis, ILF2, IL1RAP and IL1RAPL1 are the only interleukins (or interleukin receptors) that originated in invertebrates. Lineage IP294 shows the well-studied absence of IL-8 in mouse and rat [25, 26]. It is interesting to see that both species do contain the genes encoding the IL-8 receptor: IL8RA and IL8RB (IP254). Overall, the interleukin receptors seem to have originated earlier than the interleukins themselves. This evolutionary scenario of recruiting an ancient receptor into partnership with a novel ligand, seems to be proven by a recent study [27].

Table 5. Number of genes per interleukin and per species

IPID	Sc	Ce	Ag	Aa	Dm	Cs	Ci	Tn	Tr	Ol	Ga	Dr	Xt	Gg	Md	Dn	Bt	Cf	Et	La	Rn	Mm	Oc	Mm	Pt	Hs	HUGO
IP094	0	1	1	**2**	1	1	1	**2**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**3**	1	1	ILF2
IP169	0	0	0	0	0	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	IL1RAP

IP181	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	**2**	1	1	1	1	1	1	1	1	1	IL1RAPL1
IP184	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	**2**	1	1	1	1	1	1	1	1	1	IL6R
IP187	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	**2**	1	1	1	1	1	IL13RA2
IP188	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	**2**	1	1	1	1	1	IL23R
IP189	0	0	0	0	0	0	0	**3**	**4**	**2**	**4**	1	1	**3**	**2**	1	**3**	**4**	**2**	**2**	**2**	**4**	**2**	**5**	**3**	**3**	CSF2RB IL2RB IL9R
IP190	0	0	0	0	0	0	0	1	1	1	1	0	1	**2**	1	1	1	1	1	1	1	1	1	1	1	1	IL21R
IP191	0	0	0	0	0	0	0	**2**	1	1	1	**2**	1	1	1	1	1	1	1	1	1	1	1	1	0	1	LEPR
IP192	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	IL12RB2
IP193	0	0	0	0	0	0	0	**2**	1	1	1	**3**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	IL6ST
IP194	0	0	0	0	0	0	0	**2**	1	1	1	**3**	0	1	1	0	1	1	**2**	**2**	1	**2**	1	**2**	**2**	**2**	IL2RG
IP195	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	IL7R
IP214	0	0	0	0	0	0	0	1	**2**	**2**	**3**	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	IL28RA
IP220	0	0	0	0	0	0	0	1	1	1	1	1	0	**4**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	IL1R1 IL1RL1 IL1RL2
IP232	0	0	0	0	0	0	0	1	**2**	**2**	1	**2**	1	**2**	**3**	**3**	**3**	**3**	**3**	**2**	**3**	**3**	**3**	**3**	**3**	**3**	IL20
IP254	0	0	0	0	0	0	0	**3**	**3**	**4**	**4**	**3**	**2**	0	**3**	1	**3**	**4**	**3**	1	**3**	**4**	0	**4**	**4**	**4**	IL8RA IL8RB
IP261	0	0	0	0	0	0	0	**2**	**2**	**3**	**2**	1	**2**	1	**3**	1	**2**	**2**	1	1	**2**	**2**	1	1	0	1	IL4I1
IP294	0	0	0	0	0	0	0	1	1	1	1	**2**	1	0	1	1	1	1	1	1	0	0	1	1	1	1	IL8
IP348	0	0	0	0	0	0	0	1	1	1	1	**3**	1	0	1	1	1	1	1	0	**2**	1	1	1	1	1	ILF3
IP350	0	0	0	0	0	0	0	1	**2**	**2**	**2**	**2**	0	**2**	**2**	1	**2**	**2**	1	**2**	**2**	**2**	1	**2**	**2**	**2**	IL20RA IL22RA2
IP369	0	0	0	0	0	0	0	1	0	**2**	0	0	0	1	1	1	1	**2**	1	1	1	1	1	1	1	1	IL1RAPL2
IP385	0	0	0	0	0	0	0	0	1	1	1	1	1	1	**4**	1	**4**	**4**	**4**	**3**	**4**	**4**	**3**	**4**	**4**	**4**	IL1B IL1F10 IL1F5 IL1RN
IP401	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	IL1R2
IP405	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	IL4R
IP421	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	IL6
IP425	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	**2**	1	1	1	1	1	IL13RA1
IP428	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	IL22RA1
IP433	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	IL9
IP435	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	**2**	**3**	1	1	1	**2**	1	**2**	**3**	**3**	IL28A IL28B IL29
IP438	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	IL5RA
IP442	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	IL21
IP447	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	**2**	1	1	1	IL2RA
IP448	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	**2**	**2**	**2**	**2**	**5**	1	**2**	**2**	**2**	**2**	**2**	IL22 IL26
IP453	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	IL18RAP
IP481	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	IL12RB1
IP482	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	IL5
IP495	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**3**	0	**2**	1	**2**	0	**2**	**2**	**2**	**2**	**2**	**2**	IL1F6 IL1F9
IP496	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	1	0	1	1	IL27
IP499	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	1	1	1	1	1	IL24
IP515	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	IL2
IP522	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	IL7
IP523	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	0	1	0	1	IL1F8
IP531	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	IL23A
IP534	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	IL27RA
IP536	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	1	1	IL3
IP539	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	IL4
IP541	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**	1	1	0	1	0	0	1	0	1	IL1F7
IP557	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	1	IL3RA

5.4.5 Toll-like receptor pathway

We used the Toll-like receptor pathway to give an example of how pathways in different species can be mapped onto each other. Figure 5 shows this TLR pathway together with all other involved components, ultimately leading to the cell proinflammatory response (right bottom). For each of the pathway's components, the occurrence in the 26 Ensembl species was checked by using the ImmunoPhyle application. Table 6 shows that large interspecies differences can be seen in several components of the TLR pathway. For example, the lipopolysaccharide binding protein (LBP) is almost only present in vertebrates, but the nematode *C. elegans* also contains five orthologs to LBP. The jun oncogene (JUN) seems to be absent in some high vertebrates such as *G. gallus*, *D. novemcinctus*, *L. africana* and *M. mulatta*, but has expansions in the fish species (four to seven orthologs). TLR5 has multiple orthologs in the fish species as well, whereas all the other species have a maximum of only one TRL5 gene. These differences can be (part of) the solution for differences observed in immune responses between the studied species.

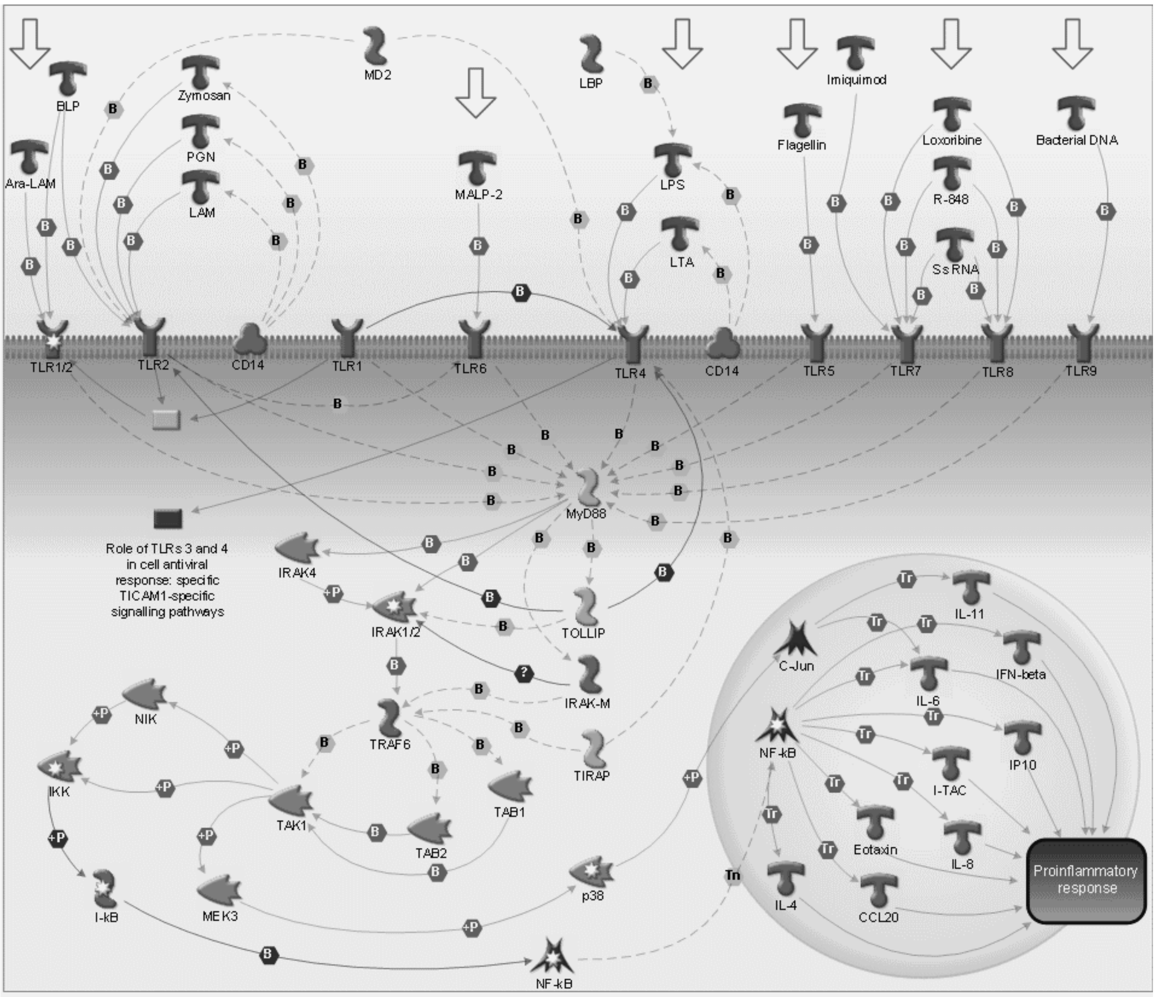


Figure 5. The toll-like receptor pathway

The pathway 'Toll-like receptor (TLR) ligands and common TLR signalling pathway leading to cell proinflammatory response' from the GeneGo MetaCore™ [28] application. *Color version on page 159.*

Table 6. Number of genes in the toll-like receptor pathway per component and per species

IPID	Sc	Ce	Ag	Aa	Dm	Cs	Ci	Tn	Tr	Ol	Ga	Dr	Xt	Gg	Md	Dn	Bt	Cf	Et	La	Rn	Mm	Oc	Mm	Pt	Hs	HUGO
IP406	0	0	0	0	0	0	0	0	0	0	0	0	**2**	1	1	**2**	**2**	**2**	0	1	**2**	**2**	**3**	**2**	**3**	**3**	TLR1/6/10
IP308	0	0	0	0	0	0	0	1	1	1	1	1	0	**2**	1	0	1	1	0	0	1	1	1	1	1	1	TLR2
IP197	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	TLR3
IP430	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	TLR4
IP289	0	0	0	0	0	0	0	**2**	**2**	**2**	**3**	1	1	1	1	0	1	1	1	0	1	0	1	0	0	1	TLR5
IP359	0	0	0	0	0	0	0	1	1	1	**2**	**4**	1	1	1	1	0	1	**2**	**2**	1	**2**	**2**	1	**2**	**2**	TLR7/8
IP550	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	1	TLR9
IP458	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	0	1	1	IRAK1
IP475	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1	IRAK2
IP397	0	0	0	0	0	0	0	0	0	0	1	**3**	0	0	1	1	1	1	1	1	1	1	1	1	1	1	IRAK3/IRAK-M
IP321	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	IRAK4
IP539	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	IL4
IP421	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	IL6
IP294	0	0	0	0	0	0	0	1	1	1	1	**2**	1	0	1	1	1	1	1	1	0	0	1	1	1	1	IL8
IP145	0	0	0	1	0	1	1	1	1	1	**2**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	MAP3K7/TAK1
IP078	0	**5**	0	0	0	1	**2**	**3**	**3**	**3**	**4**	1	**3**	**4**	**5**	**3**	**4**	**4**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	LBP
IP484	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	LTA
IP057	0	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	TOLLIP
IP045	0	1	1	1	1	**2**	**2**	**4**	**4**	**3**	**4**	**4**	**3**	**3**	**3**	1	**4**	**4**	**3**	**3**	**4**	**4**	**4**	**2**	**4**	**4**	NFKB1,NFKB2,NFKBIA
IP059	0	1	1	1	1	1	1	**7**	**7**	**5**	**6**	**4**	**4**	0	**3**	0	**3**	**2**	**2**	0	**2**	**3**	**2**	0	**2**	**3**	JUN/JUNB/JUND

5.5 Discussion

We give the first real overview of the molecular evolution of the immune system from model organisms to man. Our analysis gives general insights in this evolution and offers a framework for further investigation of interesting observations. General trends, such as the emergence of the adaptive immune system and the decline of the innate immune system, can be observed very easily. As shown in the case studies, this approach can also be used to zoom in on specific gene families or pathways. However, in order to explain differences in drug response between a certain model organism and man, usually more data is needed than just orthology data. A combination of orthology data, expression data, protein interaction data and structural data as used in recent other studies [29, 30] might help solving the problems that are encountered when transferring experimental results from model organism to man.

5.6 Acknowledgements

This work was supported financially by NV Organon. The authors thank GeneGo for their permission of using pictures from GeneGo MetaCore™.

5.7 References

1. Dempsey PW, Vaidya SA, Cheng G: The art of war: Innate and adaptive immune responses. *Cell Mol Life Sci* 2003, 60(12):2604-2621.
2. Mushegian A, Medzhitov R: Evolutionary perspective on innate immune recognition. *J Cell Biol* 2001, 155(5):705-710.
3. Kasahara M: What do the paralogous regions in the genome tell us about the origin of the adaptive immune system? *Immunol Rev* 1998, 166:159-175.
4. Flajnik MF, Du Pasquier L: Evolution of innate and adaptive immunity: can we draw a line? *Trends Immunol* 2004, 25(12):640-644.
5. Alder MN, Rogozin IB, Iyer LM, Glazko GV, Cooper MD, Pancer Z: Diversity and function of adaptive immune receptors in a jawless vertebrate. *Science* 2005, 310(5756):1970-1973.
6. Kimbrell DA, Beutler B: The evolution and genetics of innate immunity. *Nat Rev Genet* 2001, 2(4):256-267.
7. Beutler B: Innate immunity: an overview. *Mol Immunol* 2004, 40(12):845-859.
8. Pancer Z, Cooper MD: The evolution of adaptive immunity. *Annu Rev Immunol* 2006, 24:497-518.
9. Cooper MD, Alder MN: The evolution of adaptive immune systems. *Cell* 2006, 124(4):815-822.
10. Clark R, Kupper T: Old meets new: the interaction between innate and adaptive immunity. *J Invest Dermatol* 2005, 125(4):629-637.
11. Smith J, Speed D, Law AS, Glass EJ, Burt DW: In-silico identification of chicken immune-related genes. *Immunogenetics* 2004, 56(2):122-133.
12. Hutton JJ, Jegga AG, Kong S, Gupta A, Ebert C, Williams S, Katz JD, Aronow BJ: Microarray and comparative genomics-based identification of genes and gene regulatory regions of the mouse immune system. *BMC Genomics* 2004, 5(1):82.
13. Kelley J, de Bono B, Trowsdale J: IRIS: a database surveying known human immune system genes. *Genomics* 2005, 85(4):503-511.
14. Beck S: Immunogenomics: towards a digital immune system. *Novartis Found Symp* 2003, 254:223-230; discussion 230-223, 250-222.
15. Hill AV: Immunogenetics and genomics. *Lancet* 2001, 357(9273):2037-2041.
16. Miretti MM, Beck S: Immunogenomics: molecular hide and seek. *Hum Genomics* 2006, 2(4):244-251.
17. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T et al: Ensembl 2006. *Nucleic Acids Res* 2006, 34(Database issue):D556-561.
18. Hulsen T, de Vlieg J, Groenen PM: PhyloPat: phylogenetic pattern analysis of eukaryotic genes. *BMC Bioinformatics* 2006, 7:398.
19. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H: The HUGO Gene Nomenclature Committee (HGNC). *Hum Genet* 2001, 109(6):678-680.
20. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2000, 28(1):10-14.
21. Page RD: TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 1996, 12(4):357-358.

22. Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM: FlyBase: genomes by the dozen. *Nucleic Acids Res* 2006.

23. Snel B, Lehmann G, Bork P, Huynen MA: STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 2000, 28(18):3442-3444.

24. Remm M, Storm CE, Sonnhammer EL: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001, 314(5):1041-1052.

25. Modi WS, Yoshimura T: Isolation of novel GRO genes and a phylogenetic analysis of the CXC chemokine subfamily in mammals. *Mol Biol Evol* 1999, 16(2):180-193.

26. Singer M, Sansonetti PJ: IL-8 is a key chemokine regulating neutrophil recruitment in a new mouse model of Shigella-induced colitis. *J Immunol* 2004, 173(6):4197-4206.

27. Bridgham JT, Carroll SM, Thornton JW: Evolution of hormone-receptor complexity by molecular exploitation. *Science* 2006, 312(5770):97-101.

28. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T: Pathway mapping tools for analysis of high content data. *Methods Mol Biol* 2007, 356:319-350.

29. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006, 7(4):R31.

30. von Mering C, L JJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P: STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 2006.

Chapter 6

Dynamics of head-to-head genes in vertebrates

Erik Franck, Tim Hulsen, Martijn A. Huynen, Nicolette H. Lubsen, Wilfried W. de Jong and
Ole Madsen

Manuscript in preparation

6.1 Abstract

A remarkable feature of the human genome is the abundance of adjacent gene pairs (< 1kb in distance) oriented in a head-to-head (h2h) manner. This type of bidirectional gene arrangement makes overlapping and/or shared promoter elements possible, which could have biological importance. Where previous comparative analyses on the evolution of adjacent h2h genes pairs mainly focused on only h2h gene pairs and/or a limited number of genomes, we here present a comprehensive analyses of the evolution of h2h gene pairs in a variety of vertebrate genomes in relation to the evolution of the two other gene pair orientations: head-to-tail (h2t) and tail-to-tail (t2t). A random gene organization is observed in most vertebrate and invertebrate genomes in analyses without any constrains in distances between neighboring genes, whereas analyses on adjacent gene pairs (<600 bp) show an enrichment of h2h gene pairs in tetrapods (amphibians, birds and mammals), an enrichment of h2t gene pairs in Tetraodontidae fishes plus urochrotates and close to random gene organization in invertebrates, indicating a transition from a random gene organization pattern in invertebrates to an enrichment of adjacent h2t gene pairs in fish/urochordates and an enrichment of h2h gene pairs in tetrapods. While previously the border for enrichment of h2h gene pairs was set arbitrary at 1000bp, we found that the enrichment of h2h gene pairs in tetrapods is defined by an intergenic distance of less than 600bp with an optimum in human at 150bp. Analyses, in other vertebrates, on the conservation in orientation and intergenic distance of adjacent human gene pairs, shows that h2h gene pairs are not better conserved in orientation than are adjacent human h2t or t2t gene pairs, however h2h gene pairs are better conserved as adjacent gene pairs (<600 bp) than are h2t and t2t gene pairs. This indicates an evolutionary constraint against separating adjacent h2h gene pairs.

Then tracing back the origins of adjacent human gene pairs in the vertebrate lineage, it is noticeable that the interchromosomal events leading to present human adjacent gene pairs are the same for the three gene patterns, whereas intrachromosomal events leading to a decrease in intergenic distance happened earlier in vertebrate evolution for h2h gene pairs than for h2t and t2t gene pairs.

Expression analyses on human and mouse tissue data show that all three adjacent gene patterns are positively correlated when compared to random, with h2h gene pairs only slightly more co-expressed than h2t and t2t gene pairs. Furthermore, a reverse convergent relation in gene pair co-expression with increased intergenic distance is observed, which could in part count for the higher co-expression of h2h gene pairs.

6.2 Introduction

In recent years the number of sequenced vertebrate genomes has increased dramatically, making possible a variety of comparative analyses which finally will result in a profound understanding of the evolution of vertebrate genomes in general and the evolution of the human genome in particular. One notable discovery in the human genome, alongside the unexpected relative small numbers of genes, is the enrichment of gene pairs located head-to-head (h2h) with less than 1000bp between transcription start. This type of gene pair arrangement results in so-called bi-directional promoters which could have overlapping promoter activities.

Furthermore, their enrichments in the human genome indicate some kind of organized gene structure in higher eukaryote genomes in line with what is known from the prokaryote genome and could therefore be of biological importance.

Although h2h gene pairs have been known in vertebrate genomes for more than 20 years, Adachi and Lieber in 2002 [1] were the first to recognize an enrichment of bi-directional gene pairs in the human genome in analyses of chromosome 21 and 22. Later analyses on whole genomes showed that an enrichment of h2h gene pairs is also present in rodents [2,3,4] and analyses on expression patterns of human h2h gene pairs indicated either positive [2], positive and negative [4] or no [5] correlation in expression of adjacent human h2h gene pairs.

We here present a comparative analysis of adjacent h2h gene pairs in vertebrate genomes and relate their evolution to the evolution of adjacent h2t and t2t gene pairs.

6.3 Materials and Methods

6.3.1 Data sets and gene distribution

In Ensembl (Version 40) [6] the following species were used: *Homo sapiens, Pan troglodytes, Macaca mulatta Rattus norvegicus, Mus musculus, Oryctolagus cuniculus, Canis familiaris, Bos taurus, Loxodonta africana, Dasypus novemcinctus, Echinops telfairi, Monodelphis domestica, Gallus gallus, Xenopus tropicalis, Tetraodon nigroviridis, Takifugu rubripes, Danio rerio, Gasterosteus aculeatus, Ciona intestinalis, Ciona savignyi, Caenerhabditis elegans, Drosophila melanogaster, Anopheles gambiae* and *Saccharomyces cerevisiae*. For all species the number of h2h, h2t and t2t gene pairs was determined, together with the length of the intergenic region, by means of python scripting (available from the authors) on the Ensembl Ensmain genome data.

For most of the analysis we used all of the Ensembl species (data shown in supplementary data) but for practical purpose we only show and discuss the following species: Pt, Mm, Md, Gg, Xt, Tn, Dr and Ci. These species were chosen to provide a good balance between taxon sampling and quality of genome assembly, gene build and annotation.

6.3.2 Conservation and dynamics of gene pairs

To determine whether human gene pairs have a conserved localization in other species the list of human gene pairs was compared to these species via the cross species homology data (Ensortho). Of all the human gene pairs ortholog gene pairs were constructed by making combinations of the possible orthologs of the individual genes in that particular gene pair. Of each combinational ortholog gene pair the localization was determined with possible outcome h2h, h2t, t2t or ‘dc’ (different chromosomes / contigs). In case of localization at the same chromosome the intergenic distance was determined as well whether or not the two genes were separated by another gene (GI, gene insertion). Of all combinational orthologous gene pairs the most probable gene pair was chosen by taking the gene pair which most resembled the human situation, either in orientation and/or distance. For query gene pairs for which no orthologous gene pair could be found the term ‘no’ (no ortholog) was used. With this data not only the conservation of human gene pairs could be determined but it also gives an overview of the faith of these human gene pairs during vertebrate evolution. We made a vertebrate tree consisting of Hs, Mm, Rn, Gg, Xt, Tn, Tr, Dr and Ga in which we placed Mm and Rn together in a rodent supergroup and the four fish species in a fish supergroup. At each branching point and species in the vertebrate tree the relationship between the ortholog genes of human h2h/h2t/t2t gene pairs could be determined and by this way the events leading to these gene pairs in human are mapped. The events mapped are “linked”, event in which two genes are for the first time linked on the same chromosome but not in the orientation present in human. “Linked(hh/ht/tt)(igi)” is the same as “linked”, but here the two genes are in the same orientation present in human, either with (IGI, intergenic gene inclusion) or without genes between them.

6.3.3 Gene Expression

To study correlation of gene expression profiles between human and mouse gene pairs, we used an expression dataset consisting of a subset of pathologically normal human and mouse tissue samples from the Gene Logic BioExpress Database product [7]. The human dataset consists of 3,269 tissue samples in 115 tissue categories and 44,792 cDNA fragments, the mouse dataset of 859 tissue samples in 25 tissue categories and 36,701 cDNA fragments [8]. First, the Pearson correlations between the expression profiles of all cDNA fragments in the human set and all genes in the mouse set were calculated. A perfect correlation has a score of 1; a perfect anti-correlation has a score of -1. Second, the Affymetrix fragment IDs of the chip data were mapped to the Ensembl IDs used in our study. Finally, the correlation coefficients were mapped for human and mouse Ensembl h2h, h2t and t2t gene pairs and 6000 randomly assembled gene pairs as control. In the case that one Ensembl ID was mapped to multiple Affymetrix fragment IDs, the average of the multiple correlation coefficients was calculated.

6.4 Results

6.4.1 Gene Organization of Adjacent Gene Pairs

Ensembl [6] provides a comprehensive and integrated source of annotation of genome sequences and provides orthology links between genes in annotated genomes. This makes this database particularly suitable for analyzing the evolution of gene organization among different species and reasoning our used of Ensembl for our analyses. In a genome, tandem gene pairs can be organized in three different ways: 1) head-to-tail (h2t): neighboring genes are on the same strand, 2) head-to-head (h2h): genes on different strands with 5’ends toward each other, and 3) tail-to-tail (t2t): genes on different strands with 3’ends toward each other. In case of random gene distribution between neighboring genes the three patterns of gene organization should be 50%, 25% and 25%, respectively. When comparing the distributions of neighboring genes, without any limitation in distances between the two genes, such a random distribution pattern is observed in most vertebrate and invertebrate genomes (figure S1). This indicates that gene organization, on a whole, is random in vertebrate and invertebrate genomes. When plotting the number of h2h, h2t and t2t gene pairs against intergenic distance an enrichment of h2h gene pairs is observed in tetrapods from an intergenic distance of <600bp (figure 1). Thus the enrichment of h2h gene pairs is not restricted to mammals but go all the way back to the ancestor of tetrapods suggesting adjacent h2h gene pair enrichment as a more general phenomenon in vertebrates. Since we want to elucidate this enrichment we used <600bp as a border/query in subsequent analyses and not <1000bp which has been used by others in previous analyses. Analyses of adjacent gene pairs in a variety of animal species are shown in figure 2. In invertebrates and the fish species *Danio rerio* and *Gasterosteus aculeatus* a close to random gene organization of adjacent gene pairs is observed, whereas in urochordates and Tetraodontidae fishes an enrichment of h2t gene pairs and in tetrapods an enrichment of h2h gene pairs is observed. Thus the pattern in distribution of adjacent gene pairs seems to be very dynamic and a transition from random via enriched h2t to enriched h2h gene pairs seems to have occurred through the evolution of vertebrates.

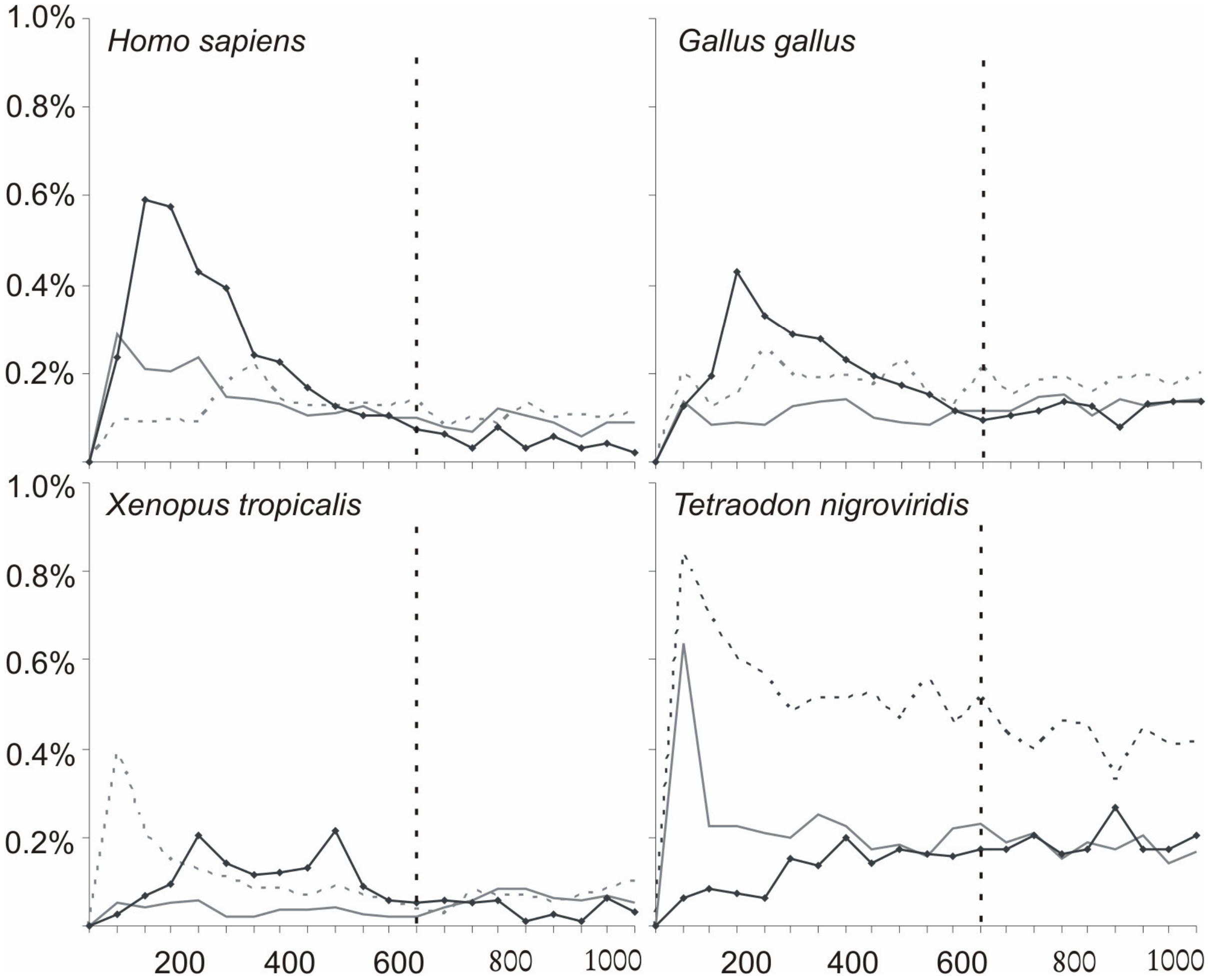


Figure 1. Frequency of adjacent head-to-head (black with diamonds), head-to-tail (gray) and tail-to-tail (dotted gray line) gene pairs in *Homo sapiens*, *Gallus gallus*, *Xenopus tropicalis* and *Tetraodon nigroviridis*. The frequency is calculated by dividing the number of gene pairs per 50bp of intergenic distance by the total number of gene pairs in the genome. The vertical dotted lines indicate the border for enrichment of H2H gene pairs.

As seen in figure 2, the described general trends vary between different species such as e.g. between the different mammals. This could illustrate the natural dynamics of the distribution of adjacent gene pairs, but incomplete genome assembly (e.g. some genomes are only sequenced with two time coverage) and gene annotation may also play a role. Comparisons between different versions of Ensembl indicate that this sometimes could be relevant. Especially between the first versions of a genome in Ensembl large changes in the results are observed, whereas after a few updates hardly any differences are found. Thus, the genomes of several query species will probably be improved in genome assembly, gene build and annotation in future Ensembl releases and hereby give a more accurate view of the differences between species. In order to avoid this problem as much as possible we choose to use the presumable best annotated genomes but still keeping a well spread sampling of vertebrate genomes (see M&M for details).

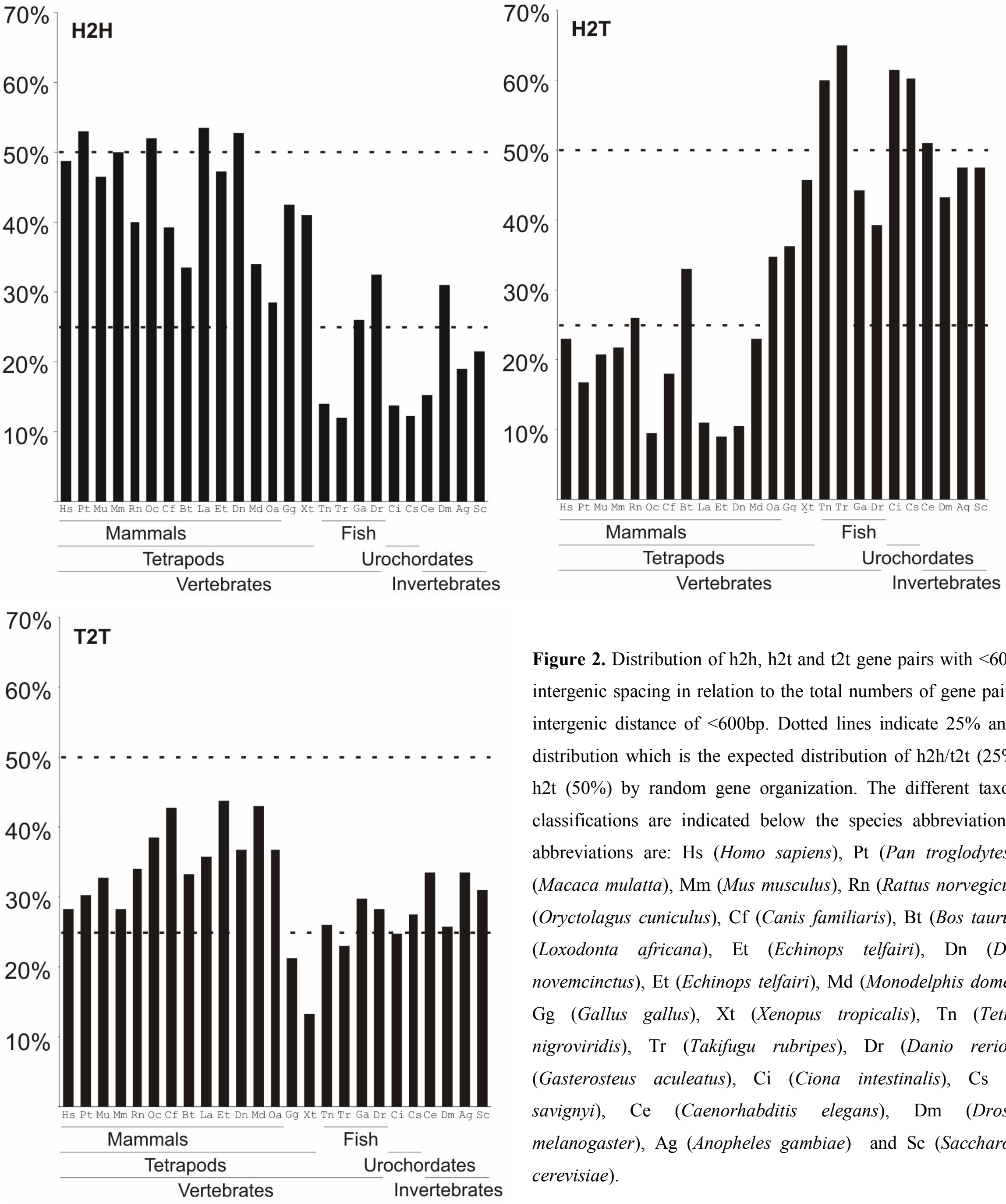


Figure 2. Distribution of h2h, h2t and t2t gene pairs with <600bp of intergenic spacing in relation to the total numbers of gene pairs with intergenic distance of <600bp. Dotted lines indicate 25% and 50% distribution which is the expected distribution of h2h/t2t (25%) and h2t (50%) by random gene organization. The different taxonomic classifications are indicated below the species abbreviations. The abbreviations are: Hs (*Homo sapiens*), Pt (*Pan troglodytes*), Mu (*Macaca mulatta*), Mm (*Mus musculus*), Rn (*Rattus norvegicus*), Oc (*Oryctolagus cuniculus*), Cf (*Canis familiaris*), Bt (*Bos taurus*), La (*Loxodonta africana*), Et (*Echinops telfairi*), Dn (*Dasypus novemcinctus*), Et (*Echinops telfairi*), Md (*Monodelphis domestica*), Gg (*Gallus gallus*), Xt (*Xenopus tropicalis*), Tn (*Tetraodon nigroviridis*), Tr (*Takifugu rubripes*), Dr (*Danio rerio*), Ga (*Gasterosteus aculeatus*), Ci (*Ciona intestinalis*), Cs (*Ciona savignyi*), Ce (*Caenorhabditis elegans*), Dm (*Drosophila melanogaster*), Ag (*Anopheles gambiae*) and Sc (*Saccharomyces cerevisiae*).

6.4.2 Mechanisms of h2h creation

A possible explanation for the enrichment of h2h gene pairs in tetrapods could be a selective elevating in gene clustering. To test this hypothesis the level of clustering was determined by counting the occurrence in which an adjacent gene pair has a neighboring gene located within 1kb (figure 3a). It appears that adjacent h2h genes have a considerable lower percentage of clustering than adjacent h2t and t2t gene pairs, indicating that gene clustering is not responsible for the enrichment of adjacent h2h gene pairs.

Another possibility for the source of adjacent h2h genes could be gene duplication events. If so the numbers of adjacent paralog h2h gene pairs should be higher than for h2t and t2t gene pairs. Paralogous genes within a species was identified via the paralogy links in Ensembl, and the percentages of paralog gene pairs within the three gene pair organizations are shown in figure 3b. Ten percent of the total group of h2t gene pairs is made up of paralog genes while in the group of h2h and t2t gene pairs four percent are made up of paralog genes. For the group of adjacent h2t and t2t genes the level of paralogous gene pairs is halved to respectively five and two percent, however the decline in adjacent paralogous h2h gene pairs is 75% (from four to one percent). Thus, adjacent h2h genes contain half as many paralogous gene pairs relative to adjacent h2t and t2t gene pairs, excluding gene duplications as a driving factor in creating adjacent h2h genes.

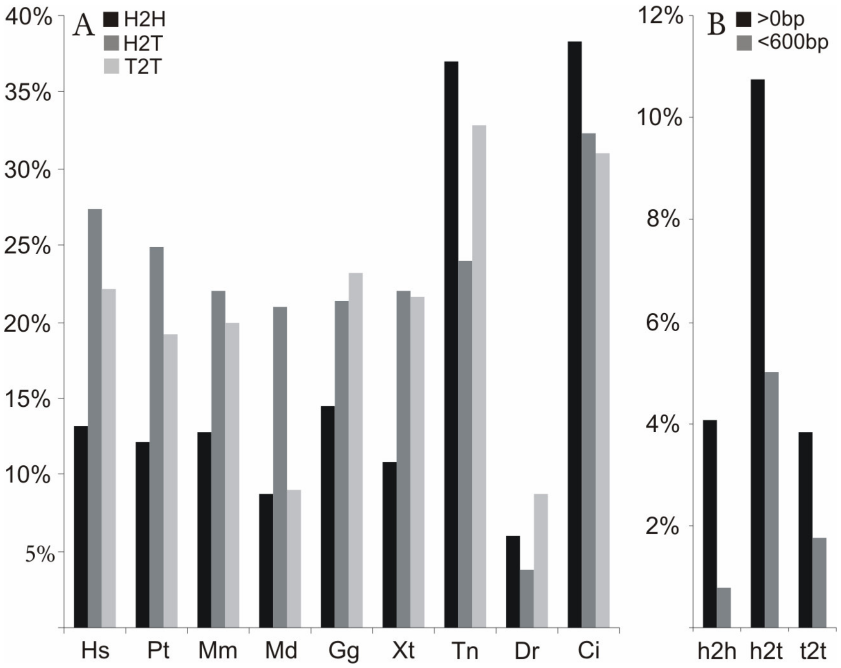


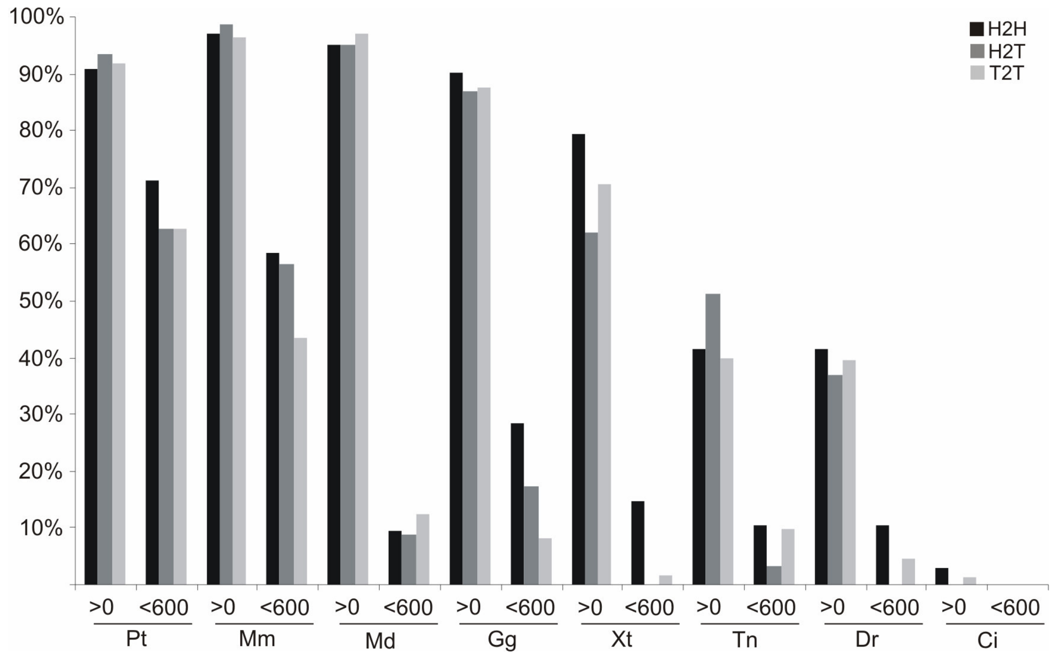
Figure 3. (a) The level that adjacent h2h, h2t, t2t gene pairs (<600bp) are part of a gene clustering. The level of gene clustering is defined by the percentage of adjacent gene pairs having a neighboring gene within 1000bp. Species abbreviations as in figure 2. **(b)** Percentage of human h2h, h2t and t2t gene pairs which consist of paralog genes.

6.4.3 Conservation and Dynamics of enriched h2h gene pairs

If there is an evolutionary pressure for adjacent h2h genes to stay together due to functional relevance than this gene arrangement should be more conserved than that of h2t or t2t genes. Therefore it would be interesting to compare the level of conservation of the adjacent h2h gene organization to that of the adjacent h2t and t2t gene organizations. For the adjacent human h2h, h2t and t2t genes the gene organization conservation was determined both by looking whether the orientation is conserved and also if the orientation and close proximity (<600bp) is conserved (figure 4a). The conservation percentages are relative to the number of orthologous gene pairs found, thus correcting for missing orthologs due to incomplete gene annotation and orthology determination. The conservation of gene organization is equal for adjacent human h2h, h2t and t2t genes, when only using the orientation criterion. When using both the orientation and distance criteria, then there is more conservation of adjacent h2h genes than h2t and t2t genes, mostly in non-mammalian species. Concluding that there is no higher evolutionary pressure for keeping adjacent h2h genes in that organization compared to h2t and t2t genes, but there is more pressure for keeping adjacent h2h gene pairs together than there is for h2t and t2t gene pairs. In addition we also looked at what organization gene pairs gained if not conserved (figure 4b). The results are in general the same for the three gene organization patterns. The greater part of non-conserved gene pairs are located on different chromosomes, but if still located on the same chromosome the majority of the changes in organizations are due to a single gene inversion (e.g. from h2h to h2t) and a lesser part to two subsequent gene inversions (e.g. from h2h to t2t).

To understand the characteristics of events leading to the enrichment of adjacent h2h gene pairs in humans the history of adjacent h2h gene pairs (and adjacent h2t/t2t for comparison) was tracked down to the divergence of chordates and vertebrates. Of each adjacent human gene pair (<600bp intergenic distance) the status of the orthologous gene pairs in vertebrates genomes was determined and events were plotted in a phylogenetic tree (figure 5). The tree consists of 5 branches containing the species (1) Hs, (2) Mm+Rn, (3) Gg, (4) Xt, and (5) Tn+Tr+Dr+Ga. Species were combined on branches to keep a more reliable and general view. In species phylogenetically close to humans, more intrachromosomal changes than interchromosomal took place, but in species phylogenetically further apart more interchromosomal changes took place than intrachromosomal changes. The most intrachromosomal events happened early in vertebrate evolution, later interchromosomal events led to adjacent gene pairs.

(a)



(b)

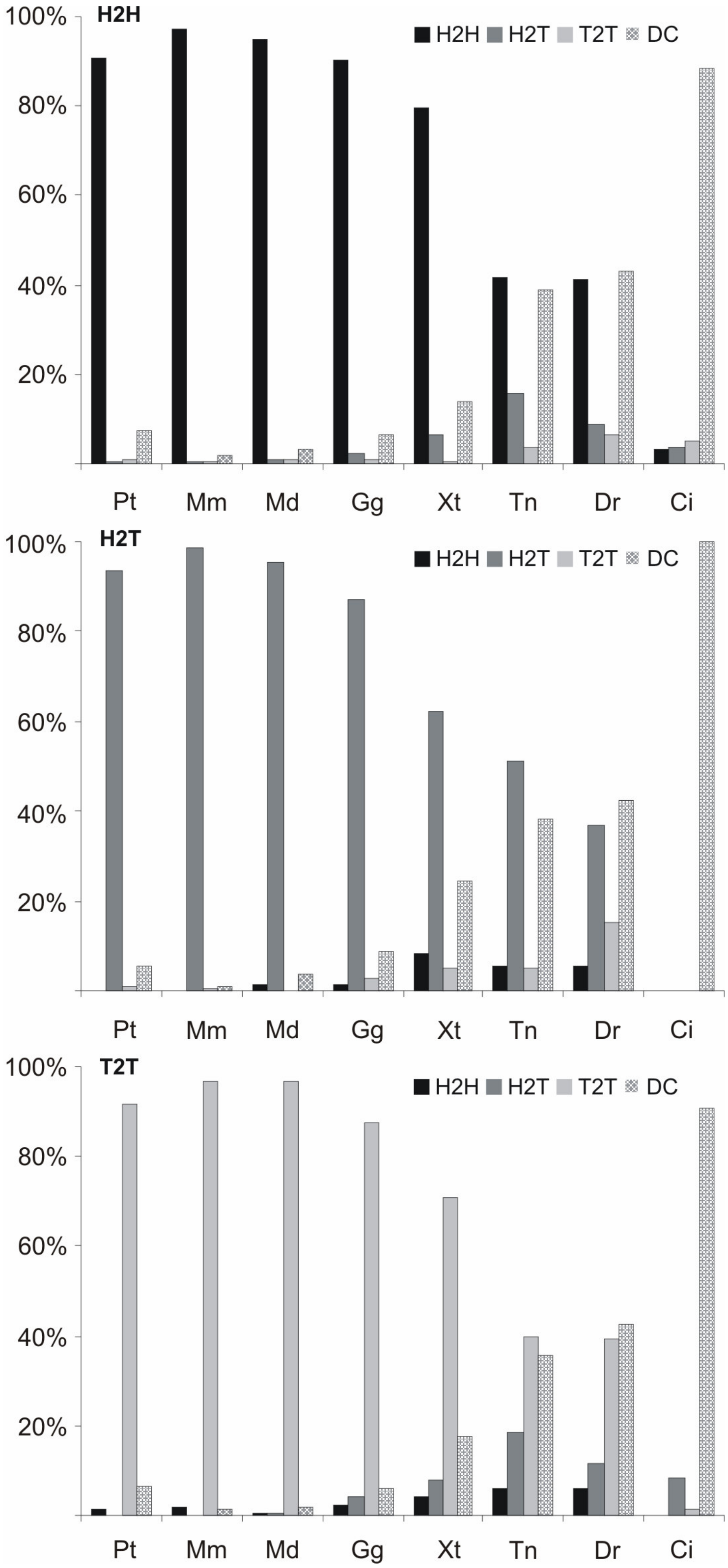


Figure 4. (a) Conservation of adjacent human h2h, h2t and t2t gene pairs. Conservation was determined by looking whether the orientation is conserved (first column) or if the orientation and close proximity (<600bp) is conserved (second column). The percentages of conservation are relative to the number of orthologous gene pairs found, thus correcting for missing orthologs due to incomplete gene annotation and/or orthology determination. **(b)** Localization of non-conserved human adjacent gene pairs. Conserved adjacent human h2h, h2t and t2t gene pairs are displayed alongside location of non-conserved gene pairs. DC indicates that the human gene pair is located on different chromosome/contigs in that species. Species abbreviations as in figure 2.

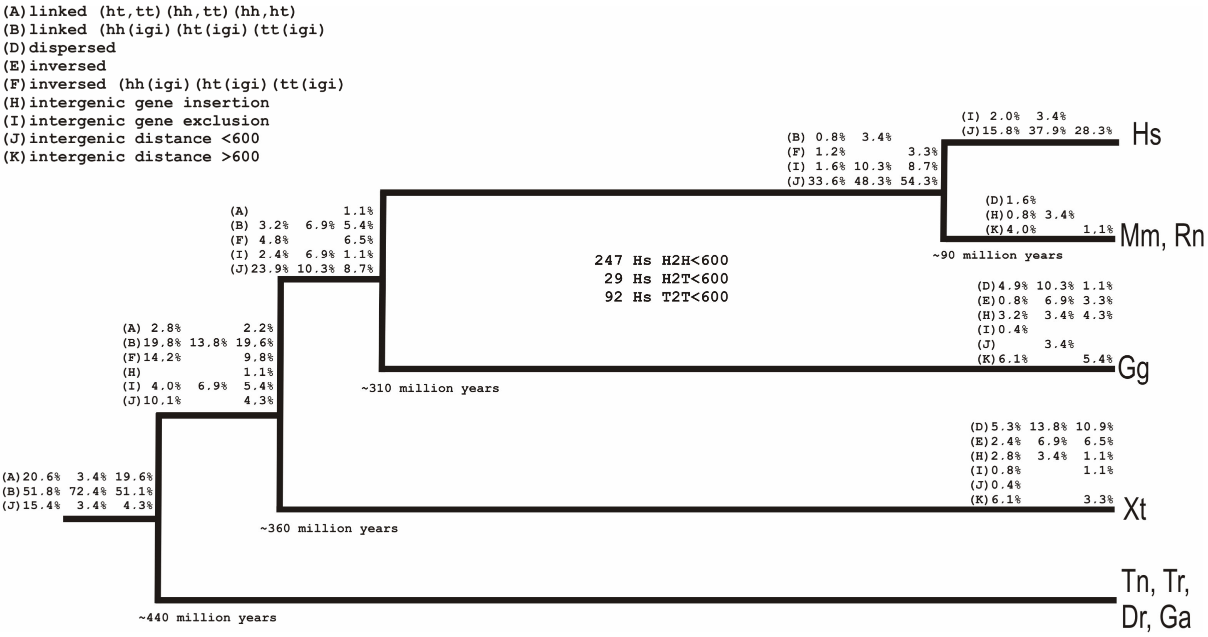


Figure 5. Of each adjacent human gene pair (<600bp intergenic distance) the status of the orthologous gene pairs in vertebrates genomes was determined and plotted in the vertebrate tree. The tree consists of 5 branches containing the species Hs, rodents (Mm, Rn), Gg, Xt and fish (Tn, Tr, Dr, Ga).

6.4.4 Expression

The transcriptional co-regulation of head-to-head promoters via potential shared cis-elements has been subject to several investigations, with similar but not equal results due to small differences in methodology and datasets. All methods used microarray datasets to construct distribution plots of the Pearson expression correlation coefficients between head-to-head genes. The general idea is that adjacent h2h gene pairs are positively or negatively co-regulated resulting in the form of Pearson correlation coefficients more closely to values of -1 (anti-regulation) or +1 (co-regulation) when comparing to non-regulated gene pairs. One report [1] mentions a shift in the distribution towards positive correlation while another [10] reports a bimodal distribution with peaks towards negative and positive correlation. Both reports use the same methodology but different cell line specific microarray datasets.

By using the Affymetrix microarray data from the Gene Express database the Pearson correlation coefficients between co-regulation and gene organization and/or intergenic distance length could be investigated for 115 human tissue categories and 25 mouse tissue categories. The tissue categories are made up of 3269 human samples and 859 mouse samples which are all non-diseased normal tissues. The correlation coefficients for the control (3249 random human gene pairs; 2197 random mouse gene pairs) and adjacent h2h, h2t and t2t gene pairs were plotted for human and mouse (figure 6a). The control plot for human follows a normal distribution, however for mouse the control plot is not normally distributed but has a slight positive skewness. In both human and mouse the plots for adjacent h2t and t2t genes are positively skewed as well but retain their mode around zero, although h2t has a large shoulder at 0.6 correlation coefficient. In contrast stand both human and mouse h2h plots which are still normally distributed but the plot have become wider and lower (platykurtic distribution) and have shifted to the positive side resulting in a mode of the plots at 0.1 and 0.2 correlation coefficient respectively. Although all three gene organizations at short distances (<600bp) become more co-regulated and less anti-regulated than the control, the effect for h2h is larger.

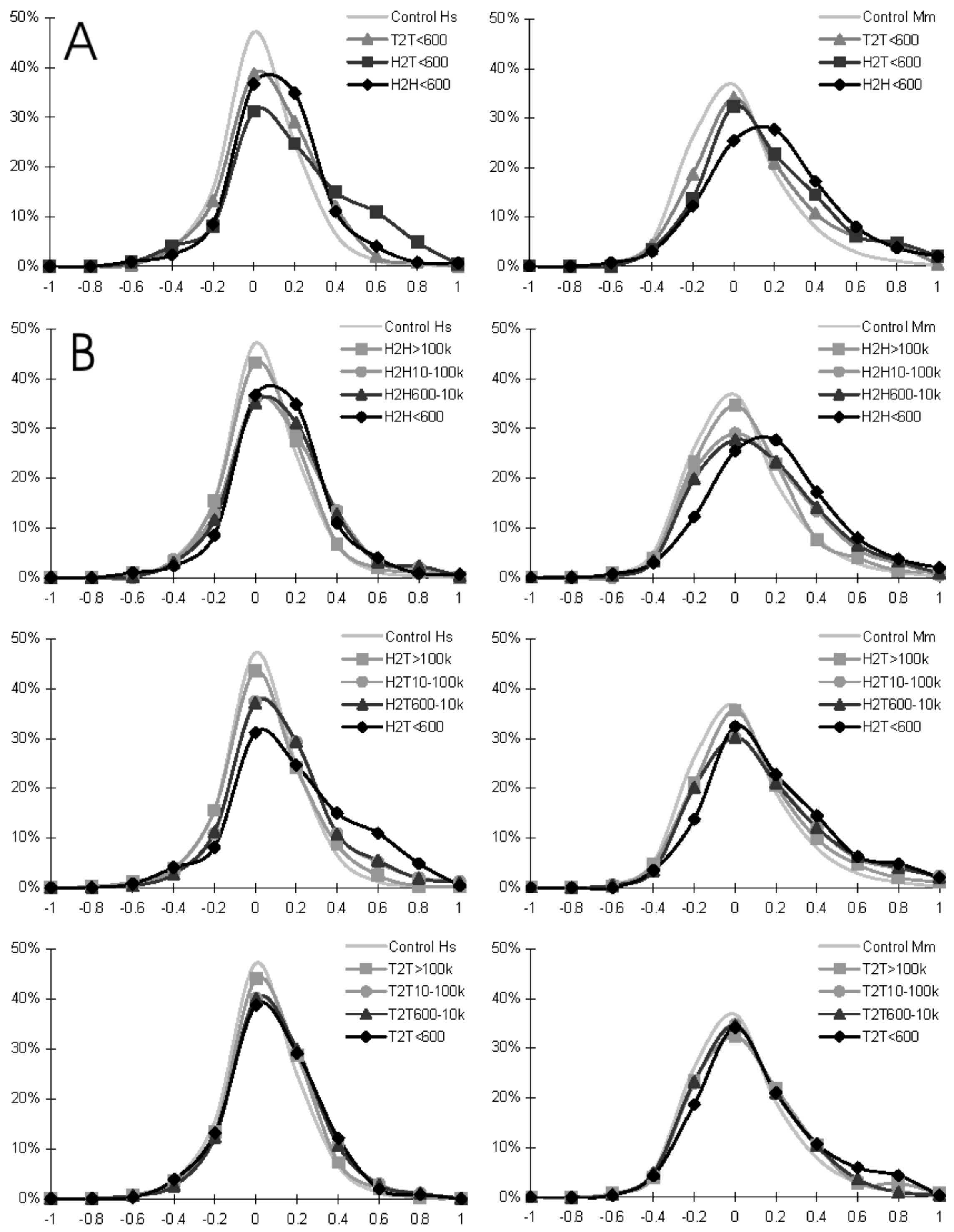


Figure 6. Expression correlation of adjacent h2h, h2t and t2t gene pairs for which microarray data was available and for a random gene pair dataset. The relative number of gene pairs with specific Pearson correlation coefficients are plotted against the Pearson correlation coefficients in increments of 0.2 units. **(A)** Human distribution plot of 599 h2h (<600bp), 247 h2t (<600bp), 364 t2t (<600bp) and 3249 random gene pairs and mouse distribution plot 401 h2h (<600bp), 145 h2t (<600bp) and 252 t2t (<600bp) and 2197 random gene pairs. **(B)** Human and mouse distribution plots of h2h, h2t and t2t gene pairs in several intergenic distance intervals (0-600bp; 600bp-10kb; 10kb-100kb and 100kb-∞).

Not only the organization of genes effects the expression correlation but also the general distance between genes is an important contributor to expression correlation. When determining the distribution plots of h2h, h2t and t2t genes classified into several intergenic distance intervals (0-600bp; 600bp-2kb; 2kb-10kb; 10kb-100kb and 100kb-∞) we can clearly conclude that with decreasing intergenic distance the plots become more platykurtic and positively skewed (figure 6b), with little differences between the three gene orientations. Only when the intergenic distance drops below 2000bp, then, in human, h2h and h2t shift more to the positive side than t2t, while in mouse only h2h shifts further to the positive side compared to h2t and t2t. For the genes with less than 600bp this effect is even larger and has been described above. This indicates that positive co-expression of genes with large intergenic distances is due to chromatin opening and/or other mechanisms acting on large distances. The effect seen at short distances (less than 2000bp) could be due to the gene organization, although no hard conclusion can be made because of the variations between human and mouse.

6.5 Discussion

In this article, we described the dynamics of bidirectional gene pairs in vertebrate genomes, and observed trends that have been described previously. However, we also found some new and interesting results:

- The distance threshold that should be used is 600bp, not 1000bp. The threshold of 1000bp used in previous studies [1] was probably chosen arbitrarily.
- We have shown the enrichment of head-to-head gene pairs not only in mammals, but also in tetrapods, which is an extension of other studies [3].
- We related the evolution of head-to-head gene pairs to the evolution of head-to-tail and tail-to-tail gene pairs.
- We tracked the origin and “mechanism of arise” of human adjacent pairs.
- The expression pattern also depends on distance, thus a fraction of the positive expression profile of head-to-head gene pairs could be explained by the distance. We also use “real” tissue expression data whereas other use cell lines.
- We observe that head-to-head gene pairs are not better conserved in orientation than are adjacent human head-to-tail or tail-to-tail gene pairs, but they are better conserved as adjacent gene pairs (<600 bp). Takai and Jones [5] found that interspersed repeats are strongly excluded from adjacent gene pairs, suggesting that these type of promoters are not allowed to be “disturbed”, resulting in the enrichments of head-to-head gene pairs in tetrapods.

Chapter 6

6.6 Acknowledgements

The authors thank Gene Logic Inc. for the use of a subset of pathologically normal human and mouse tissue samples from the Gene Logic BioExpress Database product.

6.7 References

1. Adachi N, Lieber MR: Bidirectional gene organization: a common architectural feature of the human genome. *Cell* 2002, 109(7): 807-809.
2. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM: An abundance of bidirectional promoters in the human genome. *Genome Res* 2004, 14(1): 62-66.
3. Koyanagi KO, Hagiwara M, Itoh T, Gojobori T, Imanishi T: Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene* 2005, 353(2): 169-176.
4. Li YY, Yu H, Guo ZM, Guo TQ, Tu K, Li YX, Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput Biol.* 2006, 2(7): e74.
5. Takai D, Jones PA: Origins of Bidirectional Promoters: Computational Analyses of Intergenic Distance in the Human Genome. *Molecular Biology and Evolution* 2004, 21(3): 463-467.
6. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F et al: Ensembl 2005. *Nucleic Acids Research* 2005, 33(suppl_1): D447-D453.
7. Gene Logic BioExpress Database product [http://www.genelogic.com/genomics/bioexpress/]
8. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006, 7(4): R31.

Chapter 7

The construction of genome-based transcriptional units

Sander van Hooff, Ramin Monajemi, Jan Koster, Tim Hulsen, Marco Roos, Barbera D.C. van Schaik, Marinus F. van Batenburg, Rogier Versteeg and Antoine H.C. van Kampen

Submitted

7.1 Abstract

7.1.1 Motivation

Public sequence databases typically contain many EST and (full length) mRNA sequences for every gene. For many applications it is required that all sequences that correspond to the same gene are grouped together in gene oriented sequence clusters. We present an algorithm that allows the construction of such transcriptional units.

7.1.2 Results

We constructed transcriptional units for fifteen (mammalian) organisms represented by the UCSC genome database. We discuss the results obtained for the human and mouse transcriptional units and compared our set of transcriptional units to the gene oriented sequence clusters obtained from the UniGene, ECgene and AceView approaches. These results show that the construction of gene oriented sequence clusters is still a challenging task and open for further improvement.

7.1.3 Availability

The transcriptional units for fifteen organisms are available from our web-site http://bioinfo.amc.uva.nl/HTMseq

7.2 Introduction

GenBank contains over one million full-length mRNA sequences and over thirty million expressed sequence tags (ESTs) for many organisms. Genes from, for example, human and mouse, are consequently covered by many full length mRNA and 3', 5' and random primed EST sequences. Several approaches were developed that automatically partition these sequences in gene oriented sequence clusters, where each cluster ideally contains all GenBank mRNA/EST sequences that correspond to a specific gene. The UniGene [1], TIGR Gene Indices (TGI) [2] and the Sequence Tag Alignment and Consensus Knowledgebase (STACK) [3] approaches are the most well-known but a range of other approaches and database have been developed [4-11]. The TIGR and STACK algorithms also provide consensus sequences for each sequence cluster. Splice variants detected by the TIGR method are represented in different clusters, while STACK incorporates different transcript variants in a single cluster.

Sets of gene-oriented sequence clusters have found many applications including gene discovery, expression analysis and identification of genomic expression patterns and analysis of alternatively spliced transcripts. Consequently, the construction of high quality sets of sequence clusters is vital for the correct interpretation of these applications.

From the methods available to generate gene-oriented sequence clusters, only UniGene [1], ECgene [10], AceView [12] and RIKEN [11] use the genomic sequence as a template for the clustering process. UniGene clusters are available for a large range of (mammalian) organisms, while ECgene only provides *H. sapiens*, *M. musculus* and *R. norvegicus* clusters. AceView currently provides clusters for *H. sapiens*, *C. elegans* and *A. thaliana* but has not been updated since 2005. The UCSC genome database [13,14] uses the genome to determine 'gene boundaries' but does not provide an explicit set of sequence clusters. Moreover, the construction of these gene boundaries is not very well documented and these gene boundaries are not updated regularly. The RIKEN set of clusters for human and mouse was also assembled using genomic information, but the resulting clusters do not include all GenBank mRNA/EST sequences. Therefore, in this study we compare our set of transcriptional units only to UniGene, ECgene and AceView.

The alignment of transcripts to the genomic sequence provides information about the genomic location of the corresponding genes and, consequently, whether a set of sequences belongs to the same gene. In addition, these alignments directly provide insight in the gene structure. Despite aforementioned efforts to generate reliable sequence clusters, it turns out that this is still a challenging task that remains open for improvement.

The UniGene set [1] of gene oriented clusters is probably the most frequently used set but not all details are documented. UniGene uses a genome-based clustering approach to identify sets of transcript sequences that correspond to individual genes.

The program 'Splign' (http://www.ncbi.nlm.nih.gov/sutils/splign) is used to align transcripts against the genome. Subsequently, it records the annotated exon boundaries and the association of exons with genes. Any stringently aligned sequences that share exon-intron boundaries that are identified with only one gene are grouped together. Unspliced sequences, as well as sequences for which the splicing location or orientation is uncertain, are associated with an overlapping exon if one exists, or placed against the genome if not. Sequence orientation is used where there is possible ambiguity of gene orientation. Sequences that do not align to genomic sequence are grouped together, and transcribed sequences within an interval smaller than 3000 nucleotides that have a common clone of origin are grouped together. Clusters that do not correspond to an annotated gene and are less than 500 bases downstream 3' of another cluster are considered alternative 3' termini, and are merged into the upstream cluster.

ECgene [10,15] was developed to detect alternative splicing and starts from BLAT alignments of sequences against the genome. Erroneous BLAT alignments are corrected and, subsequently, sequences that share splice sites are grouped together. The direction of each cluster is determined from the intron consensus sequence and the presence of a polyadenylation tail. Subsequently, unspliced sequences are added to existing clusters or form new clusters. Finally neighbouring genes within 2 Mbp are merged if they contain ESTs sharing a common clone of origin.

The AceView approach uses an expert-supervised automatic annotation and is, therefore, not suitable to quickly build clusters for a large number of organisms. The AceView algorithm is not described in full detail. AceView first aligns all publicly available mRNAs and EST sequences on the genome by using an unpublished algorithm. From these alignments transcripts are reconstructed which are subsequently clustered into genes based on overlap and shared intron boundaries. Each reconstructed transcript represents a different splice form of that gene. For each reconstructed transcript a consensus sequence is determined.

Several of our research projects require the availability of a high quality set of sequence clusters but also the flexibility to change the underlying algorithms when problems are identified. Projects in which we currently use the transcriptional units include the construction of transcriptome maps [16,17], the identification of SAGE tags [16], the annotation of microarray probes and the inspection of gene structure with Transcript View [18]. This prompted for the development of the algorithms presented in this paper and which we here describe in detail. Although the overall cluster procedure is similar to the procedures taken by UniGene, AceView and ECgene, many (details of the) steps in our algorithm differ from these other approaches. We applied our algorithms to construct transcriptional units for fifteen organisms but mainly discuss the results obtained for human and mouse.

7.3 Methods

7.3.1 Construction of transcriptional units

The construction of transcriptional units comprises six steps (figure 1):

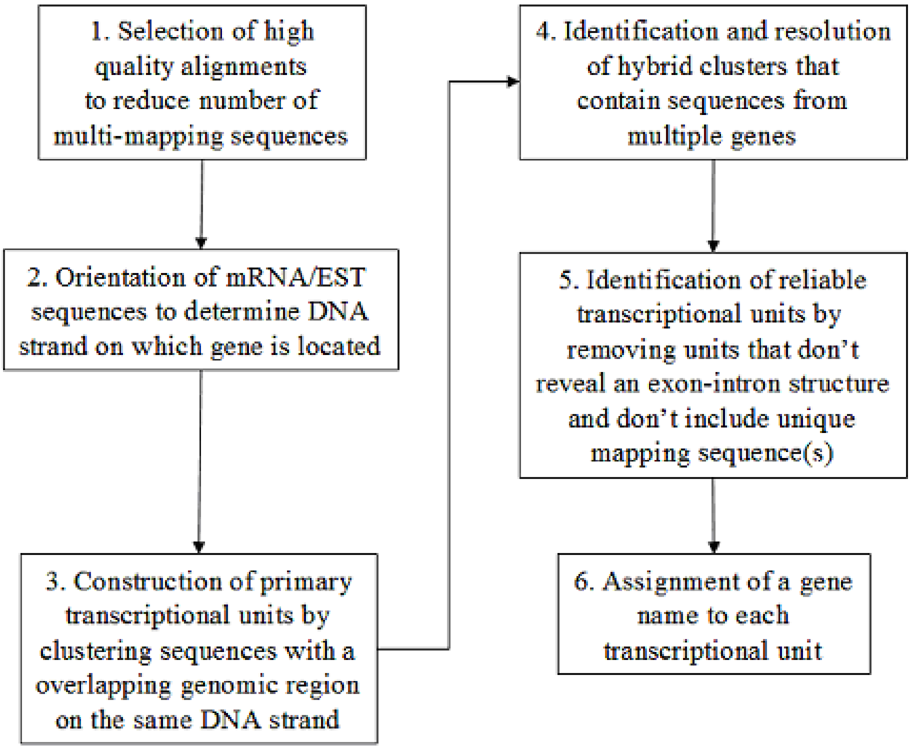

1. Selection of high quality alignments to reduce number of multi-mapping sequences
2. Orientation of mRNA/EST sequences to determine DNA strand on which gene is located
3. Construction of primary transcriptional units by clustering sequences with a overlapping genomic region on the same DNA strand
4. Identification and resolution of hybrid clusters that contain sequences from multiple genes
5. Identification of reliable transcriptional units by removing units that don't reveal an exon-intron structure and don't include unique mapping sequence(s)
6. Assignment of a gene name to each transcriptional unit

Figure 1. Flow diagram depicting the steps taken in our algorithm to construct transcriptional units.

1. *Selection of high quality alignments.* The construction of transcriptional units starts with BLAT alignments of mRNA/EST against the genomic sequence as provided by the UCSC genome database [13,14]. The UCSC genome database only supplies alignments for which the base identity is at least 96% and within 0.5% of the best alignment. Due to the presence of pseudo-genes, gene families or repeat sequences, ESTs/mRNA can map to multiple genomic positions. For UCSC build hg18 about 5.1% (374,642) of the sequences in the genome database align to multiple genome locations involving 12.1% (944,340) of the alignments. To reduce the number of multi-mapping sequences we identified the most likely genome position by selecting the alignment with the highest UCSC BLAT alignment score (which is based on the number of (mis)matches and the number of bases that align to repeat regions) and, if present, additional alignments within 3% of this score. For human, this approach reduces the number of sequences that align to multiple positions to 3.7% (276,088) involving 9.8% (731,289) of the alignments.

2. *Orientation of mRNA/EST sequences.* To prevent construction of erroneous transcriptional units (genes) that contain sequences from two overlapping genes on opposite strands one must determine the orientation of the individual sequences such that these can be assigned to the correct DNA strand. Information about the splicing sites is used to orient the sequence. The intron splice site GT/AG is used by more than 98% of all genes [19] and provides the most important source of information for orientation of sequences. Following the choice made by the UCSC genome browser, we set the minimum length of introns to 32 nucleotides in our algorithm.

If no splicing sites are identified then information about the polyadenylation tail and signal of each transcript may be used to determine the 3'-end of a sequence [20]. For our algorithm we consider the two most common signals (AATAAA and ATTAAA) which are used by 80-90% of the transcripts [21]. Several other polyadenylation signals are known [21-25] and although inclusion of these signals may increase the sensitivity for detecting 3' sequence ends, this also reduces the specificity and will, consequently, introduce many more erroneous orientation assignments [26]. Polyadenylation signals generally occur within the last 50 nucleotides of a transcript (excluding the polyadenylation tail)[27] and, therefore, we search within this region for the aforementioned two signals.

Two parameters play a role in the determination of sequence orientation from polyadenylation information. The first one is the minimum length of the polyadenylation (or polyT) stretch required to positively determine sequence orientation. The second parameter is the presence of the polyadenylation signal within the last 50 nucleotides. To determine the sequence orientation parameters we distinguish between two classes of sequences (3'-EST/mRNA/RefSeq and 5'-EST/random-EST/unknown), which are annotated by the GenBank sequence label.

To determine the optimal parameterization we constructed two training sets comprising the 3'-EST/mRNA/RefSeq sequences (N_1=268,744) and the 5'-EST/random-EST/unknown sequences (N_2=1,934,284) for which we could assign the orientation based on the splice sites and which we assumed to be correct. For these sequences we again determined sequence orientation by requiring a minimum length of the polyadenylation stretch (from L=0 to L>=19) and including information about the presence (or absence) of a polyadenylation signal. Subsequently, we compared these orientation assignments to those based on the splice sites. From these comparisons we calculated the ratio R of correct to incorrect assignments ($R = \frac{\#\text{correct orientation assignments}}{\#\text{incorrect orientation assignments}}$) and determined the fraction of sequences in these two training sets that are correctly oriented for different thresholds for the minimum polyadenylation tail length L ($fraction = \frac{\#\text{correct orientation assignments}}{\#\text{sequences in training set}}$). This fraction will decrease for thresholds that require larger polyadenylation tails since sequences with longer polyadenylation tails will be less common. At the same time longer tails will result in more reliable orientations. Consequently, there is a trade-off between the fraction and ratio R for which an optimum must be determined.

Table 1. Cumulative fraction of correctly oriented sequences from training set and ratio R for the 3' and 5' sequence classes and for different combinations of polyadenylation tail length and presence of signal.

Minimum length polyA tail	3' EST/mRNA/RefSeq				5' EST/Random			
	PolyA tail only		PolyA tail + Signal		PolyA tail only		PolyA tail + Signal	
	Fraction	R	Fraction	R	Fraction	R	Fraction	R
0	n.a.	n.a.	0.515	82	n.a.	n.a.	0.105	16
1	0.115	2	0.330	269	0.151	3	0.047	82
2	0.087	5	0.304	353	0.058	5	0.041	223
3	0.077	9	0.286	377	0.032	10	0.039	401
4	0.070	11	0.267	382	0.022	18	0.038	564
5	0.065	12	0.253	400	0.017	28	0.037	643
6	0.061	12	0.244	395	0.015	37	0.036	696
7	0.059	12	0.235	386	0.014	47	0.035	740
8	0.056	12	0.227	391	0.013	56	0.034	747
9	0.054	12	0.219	392	0.012	63	0.033	789
10	0.051	12	0.211	391	0.012	70	0.032	769
11	0.048	13	0.203	404	0.011	80	0.030	785
12	0.046	13	0.196	405	0.011	88	0.030	773
13	0.043	14	0.187	393	0.011	92	0.029	778
14	0.040	16	0.175	386	0.010	99	0.027	780
15	0.036	17	0.159	364	0.010	105	0.026	763
16	0.029	17	0.135	321	0.009	115	0.023	752
17	0.023	17	0.106	280	0.008	115	0.020	722
18	0.012	11	0.050	149	0.007	120	0.018	658
>=19	0.004	32	0.017	1545	0.006	118	0.014	603

The results of the sequence orientation are shown in table 1 and figure 2. Figure 2a shows the ratio R versus the fraction of correct orientations for 3'-EST/mRNA/RefSeq sequences. Clearly, orientation on basis of a polyadenylation tail in absence of a polyadenylation signal results in much lower R values. Figure 2b shows the results for 5'EST/random primed or unlabeled sequences. In contrast to the previous case a much lower fraction of sequences is correctly oriented. We observe that reasonable values for R (>100) are first obtained for orientation rules that also require the presence of a polyadenylation signal and tail.

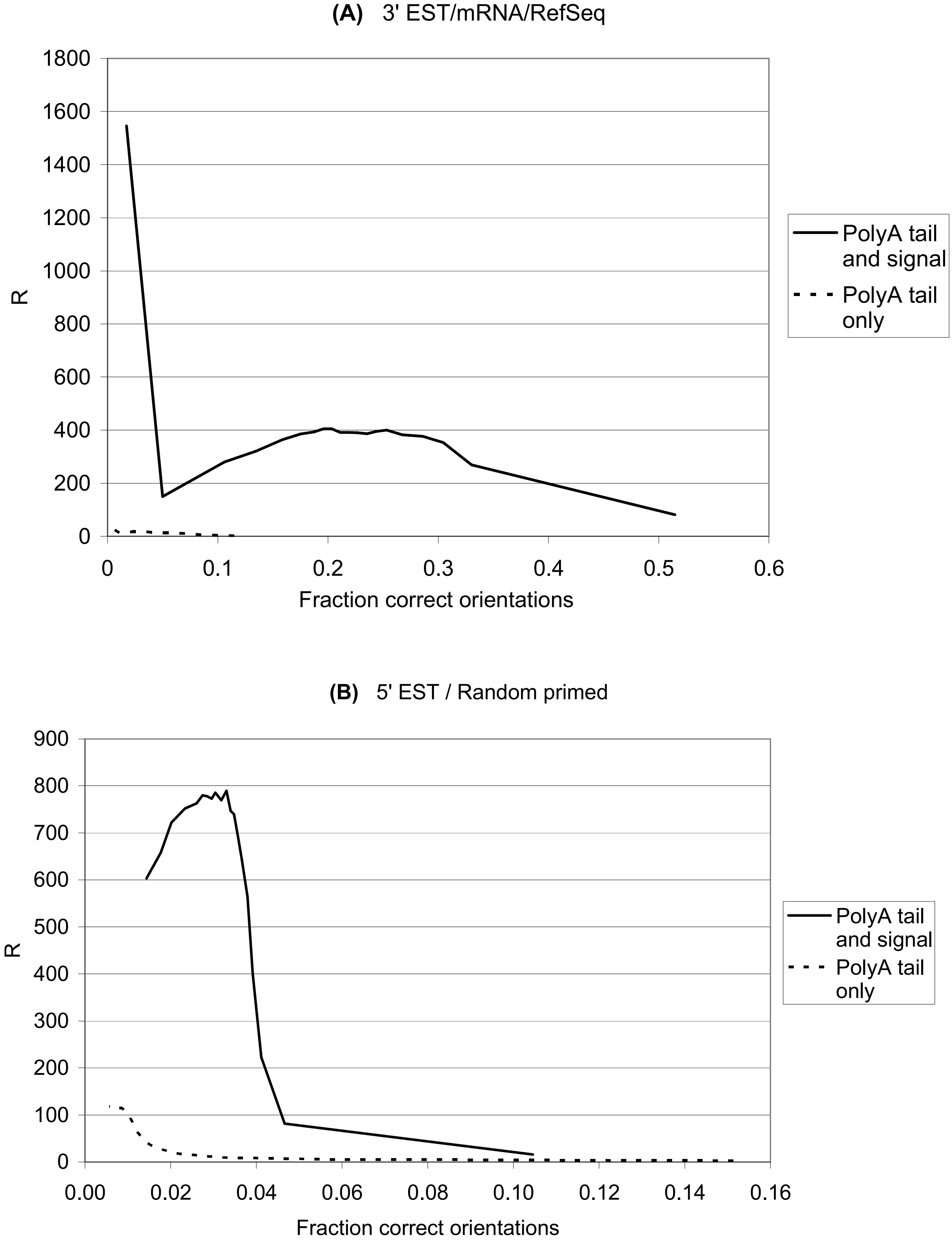


Figure 2. Different configurations of two single sequence orientation rules. R denotes the ratio of correct to incorrect orientations. The x-axis denotes the fraction of correctly oriented sequences. The different points on the curves correspond to polyadenylation tails with minimum lengths varying from >=19 to 0 (see Table 1) corresponding to small to larger fractions. The upper (blue) curve corresponds to the orientation rule in which both a polyadenylation tail and signal is required. The lower (purple) curve corresponds to the orientation rule in which only a polyadenylation tail is required. **(a)** Orientation of 3' EST/mRNA/RefSeq sequences. **(b)** Orientation of 5' EST/Random primed sequences. The curves show the trade-off between the fraction of correctly oriented sequences and ratio R. Furthermore, for different choice of the polyadenylation tail length it shows the difference between using a rule with and without polyadenylation signal.

Using the information from table 1 we can construct orientation rules for each sequence class. Potentially we can construct two rules for each sequence class, i.e., a rule for sequences that contain a polyadenylation tail and signal and a second rule (with a different parameterization) for the orientation of sequences with only a polyadenylation tail. For both rules we need to determine the optimal tail length. Subsequently, we determine whether application of the first rule or a combination of both rules gives the highest performance with respect to the fraction of correctly oriented sequences and ratio R. The optimal rule(s) can easily be found by determining the fraction and ratio R for every combination of rules. Table 2 shows the results for the 3' sequence class for the combination of two rules with different choices of the polyadenylation tail length. The yellow, blue and green colours indicate ratios with values above 50, 100 and 200 respectively. We required a ratio of at least R=100, i.e., one in every hundred assignments will be incorrect. The optimal choice of two rules is indicated by the red box and was found by identifying the highest fraction for the required minimum ratio. A first rule for the orientation of 3'-EST/mRNA/Refseq sequences that contain a signal require a minimum tail length of L=1. A second rule for 3'-EST/mRNA/RefSeq sequences that do not contain a signal requires a tail length of at least L=15. This combination of rules result in a fraction of correctly oriented sequences of 0.37 and a ratio R=109. If we, however, compare these values to the fraction (0.33) and ratio (R=269) from Table 1 that are obtained if a single rule would be selected (i.e., only orientation of sequences that contain polyA signal) we see that the relative increase in fraction is only 9% while the decrease in ratio R is 55%. We find this trade-off unfavourable and, consequently, we only use a single rule for the orientation of 3' sequences. Common choices for the tail length are L=5 or L=10 but we can now see from table 1 that although this leads to less errors in the orientation, it also significantly decreases the number of orientations assignments that can be made.

For the 5' sequence we compiled a table similar to table 2 (supplementary table S3). This resulted in the selection of two rules. The first rule requires the presence of a polyadenylation signal and a minimum tail length of L=2 and the second rule that orients the sequences without a signal requires a tail length of at least L=19. These two rules orient 5% of the 5' sequences correctly resulting in a ratio R=201. In this case the selection of two rules compared favourably against the selection of a single rule (i.e., the relative decrease of the ratio was 10% while the increase in fraction was 14%).

Table 2. Fraction correctly oriented sequences and ratio R for the 3' sequence class for different combinations of rules (polyadenylation tail+signal and polyadenylation tail only).

		PolyA signal + tail																			
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Fraction polyA tail only	**1**	0.63	0.45	0.42	0.40	0.38	0.37	0.36	0.35	0.34	0.33	0.33	0.32	0.31	0.30	0.29	0.27	0.25	0.22	0.16	0.13
	2	0.60	0.42	0.39	0.37	0.35	0.34	0.33	0.32	0.31	0.31	0.30	0.29	0.28	0.27	0.26	0.25	0.22	0.19	0.14	0.10
	3	0.59	0.41	0.38	0.36	0.34	0.33	0.32	0.31	0.30	0.30	0.29	0.28	0.27	0.26	0.25	0.24	0.21	0.18	0.13	0.09
	4	0.58	0.40	0.37	0.36	0.34	0.32	0.31	0.31	0.30	0.29	0.28	0.27	0.27	0.26	0.25	0.23	0.20	0.18	0.12	0.09
	5	0.58	0.40	0.37	0.35	0.33	0.32	0.31	0.30	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.22	0.20	0.17	0.11	0.08
	6	0.58	0.39	0.37	0.35	0.33	0.31	0.31	0.30	0.29	0.28	0.27	0.26	0.26	0.25	0.24	0.22	0.20	0.17	0.11	0.08
	7	0.57	0.39	0.36	0.34	0.33	0.31	0.30	0.29	0.29	0.28	0.27	0.26	0.25	0.25	0.23	0.22	0.19	0.16	0.11	0.08
	8	0.57	0.39	0.36	0.34	0.32	0.31	0.30	0.29	0.28	0.27	0.27	0.26	0.25	0.24	0.23	0.21	0.19	0.16	0.11	0.07
	9	0.57	0.38	0.36	0.34	0.32	0.31	0.30	0.29	0.28	0.27	0.26	0.26	0.25	0.24	0.23	0.21	0.19	0.16	0.10	0.07
	10	0.57	0.38	0.36	0.34	0.32	0.30	0.29	0.29	0.28	0.27	0.26	0.25	0.25	0.24	0.23	0.21	0.19	0.16	0.10	0.07

	11	0.56	0.38	0.35	0.33	0.32	0.30	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.24	0.22	0.21	0.18	0.15	0.10	0.07
	12	0.56	0.38	0.35	0.33	0.31	0.30	0.29	0.28	0.27	0.26	0.26	0.25	0.24	0.23	0.22	0.20	0.18	0.15	0.10	0.06
	13	0.56	0.37	0.35	0.33	0.31	0.30	0.29	0.28	0.27	0.26	0.25	0.25	0.24	0.23	0.22	0.20	0.18	0.15	0.09	0.06
	14	0.55	0.37	0.34	0.33	0.31	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.24	0.23	0.21	0.20	0.17	0.15	0.09	0.06
	15	0.55	0.37	0.34	0.32	0.30	0.29	0.28	0.27	0.26	0.25	0.25	0.24	0.23	0.22	0.21	0.19	0.17	0.14	0.09	0.05
	16	0.54	0.36	0.33	0.32	0.30	0.28	0.27	0.26	0.26	0.25	0.24	0.23	0.23	0.22	0.20	0.19	0.16	0.14	0.08	0.05
	17	0.54	0.35	0.33	0.31	0.29	0.28	0.27	0.26	0.25	0.24	0.23	0.23	0.22	0.21	0.20	0.18	0.16	0.13	0.07	0.04
	18	0.53	0.34	0.32	0.30	0.28	0.26	0.26	0.25	0.24	0.23	0.22	0.21	0.21	0.20	0.19	0.17	0.15	0.12	0.06	0.03
	19	0.52	0.33	0.31	0.29	0.27	0.26	0.25	0.24	0.23	0.22	0.21	0.21	0.20	0.19	0.18	0.16	0.14	0.11	0.05	0.02
Ratio R polyA tail only	**1**	9	7	7	6	6	6	6	6	5	5	5	5	5	5	5	4	4	4	3	2
	2	26	24	23	22	21	20	19	19	18	18	17	17	17	16	15	14	13	11	8	6
	3	39	40	39	37	36	34	33	33	32	31	30	30	29	28	27	25	23	20	14	11
	4	46	52	51	49	47	45	44	43	42	41	40	39	38	37	35	33	30	26	18	13
	5	49	58	57	55	53	51	49	48	47	46	45	44	43	41	39	37	33	28	19	15
	6	50	61	60	58	55	54	52	51	50	48	47	46	45	44	42	39	35	30	20	15
	7	51	63	63	61	58	56	55	53	52	51	49	48	47	46	43	41	36	31	21	15
	8	52	66	65	63	61	59	57	55	54	53	52	50	49	48	45	42	38	32	21	16
	9	53	68	68	66	63	61	59	58	56	55	53	52	51	49	47	44	39	33	22	16
	10	54	71	71	69	66	64	62	61	59	58	56	55	54	52	49	46	41	35	23	17
	11	56	76	76	74	70	68	66	65	63	62	60	59	57	55	53	49	44	37	24	17
	12	58	81	82	80	76	74	72	70	68	67	65	64	62	60	57	53	47	40	26	18
	13	60	89	90	88	84	82	80	77	76	74	72	71	69	67	63	59	52	44	28	20
	14	63	98	101	98	95	92	90	87	85	83	81	80	78	75	72	66	59	50	31	22
	15	65	109	113	111	107	104	101	98	96	94	92	90	88	85	81	75	66	56	35	25
	16	68	120	127	125	121	118	115	112	109	107	104	103	100	97	92	85	75	63	38	26
	17	70	138	149	147	142	140	136	133	130	127	125	123	120	116	110	102	90	75	44	30
	18	71	148	162	162	157	154	150	146	143	140	137	136	133	128	122	112	97	81	44	26
	19	81	249	316	332	334	345	339	331	333	332	330	337	335	325	316	296	260	223	120	170

The decisions made to determine sequence orientation are summarized in table 3. Inspection of the UCSC genome alignments revealed that only about 52% and 44% of the database sequences for human and mouse respectively contain a splicing site. The other sequences may represent single exons (e.g. an unspliced sequence) or use alternative donor and acceptor sites [19]. In these cases we attempt to determine the orientation of the sequence from information provided by the polyadenylation signal and tail (only 4% and 2% for human and mouse sequences respectively). All sequences that cannot be oriented are assigned to the category of 'unoriented' sequences.

Table 3. The number and percentage of sequences that are oriented by the different criteria in our algorithm.

Criteria		**Number of sequences** **(*H. sapiens*)**	**Number of sequences** **(*M. musculus*)**
Presence of exon-intron splice site		4,069,141 (52.4%)	2,054,604 (44.2%)
3'-end sequence label OR mRNA/RefSeq sequence	polyA + signal	316,094 (4.1%)	78,747 (1.7%)
5'-end, random primed sequence label or label unknown	polyA + signal	47,627 (0.6%)	22,195 (0.5%)
	polyA	4,833 (0.11%)	6,890 (0.11%)

3. *Construction of primary transcriptional units.* Overlapping sequences on the same strand and overlapping sequences in the "unoriented" category are clustered to form primary transcriptional units. We consider sequences overlapping if they share a genomic region between the start and stop position of their respective alignments, i.e., we do not require exon overlap or sharing of splice sites. Requiring exon overlap would, in principle, allow the detection of nested genes on the same strand but only few of these nested genes exist [28]. Moreover, requiring overlap in our algorithm resulted in many 'nested' transcriptional units that were artefacts caused by small EST sequences that did not have exon overlap with their neighbouring sequences. To reduce the computational complexity we therefore did not require exon overlap. This step of the algorithm results in transcriptional units, which contains all spliced and unspliced sequences for a specific gene including transcripts variants.

4. *Identification and removal of hybrid clusters.* Due to the presence of sequence and alignment artifacts the set of primary transcriptional units contain erroneous hybrid clusters that include sequences from two or more neighboring genes, which were linked by, for example, a genomic DNA or a chimeric sequence [29]. To identify and resolve hybrid clusters we first select all clusters that contain at least 25 sequences in order to have sufficient experimental evidence to confidently resolve the neighboring genes (smaller clusters are retained but not subjected to this step). Subsequently, the co-occurrence of every exon pair represented by the sequences of a transcriptional unit is counted. The sequence(s) in which the pair consisting of the first and last exon has a co-occurrence score smaller than 1 + (cluster size / 500) are removed and the remaining sequences are re-clustered to obtain the new transcriptional units. This last threshold was determined empirically from manual inspection of a test set of transcriptional units.

5. *Identification of reliable clusters.* The transcriptional units that result from the previous four steps of our algorithm may still contain sequences that align to multiple positions on the genome. This may be caused by incorrect BLAT alignments, the presence of repeat regions, the presence of pseudo-genes, and highly similar gene family members. To account for these situations and to obtain a set of reliable clusters we only retain the units that contain at least one RefSeq/mRNA sequence that maps on a single location or units that contain at least two ESTs which map on a single location and that cover at least 25% of the total exon length of the cluster to ensure that these ESTs really correspond to gene represented by the transcriptional unit. The first condition ensures the removal of transcriptional units that only contain sequences that map on multiple positions and of which the location is therefore uncertain. The second condition avoids the inclusion of multi-mapping units that would not be removed due to relatively small, and therefore possibly erroneous, alignment of an EST. However, to avoid the loss of transcriptional units that align to multiple positions but have an exon-intron structure for at least one location and, therefore, are likely to represent true genes, we retain all units that contain at least one spliced RefSeq/mRNA or two spliced ESTs that cover at least 25% over the total exon length in the alignment. This allows distinguishing between true genes and processed pseudogenes. Of the 389,117 human clusters that result from step 4 and 5 33,007 are classified as reliable (for mouse these numbers are respectively 226,059 and 27,792).

6. *Assignment of transcriptional unit names.* In this last step we assign, if possible, gene names (Gene ID's and symbols) to each transcriptional unit on basis of the accession codes of the sequences. For this step we use the Entrez "Gene" database [30]. From the 33,007 human and 27,792 mouse transcriptional units we could uniquely link 18,082 human and 20,151 mouse units to one Gene identifier. 896 human and 720 mouse units are linked to multiple entries of the Gene database. In this situation multiple names are assigned to a transcriptional unit. A large number of units (14,029 and 6,921 for human and mouse respectively) could not be linked to the Gene database. In this case we annotate the clusters either as "unknown" if the unit contains a mRNA or RefSeq sequence, or as "EST" when the unit only consists of ESTs.

7.3.2 Linking transcriptional units from different organisms

In several of our applications we require the linkage of transcriptional units for different organisms. Therefore, we identified orthologous relationships between the transcriptional units of all fifteen analyzed species. Orthologous genes were identified by using the Best Bidirectional Hit method, which is proven to be stringent and reliable [31]. Briefly, we first created FASTA files of all consensus sequences from each transcriptional unit for each species. Secondly, we performed all-against-all sequence comparisons between each pair of FASTA files. We used the BLAST2 algorithm of the Biofacet package (http://www.gene-it.com) for the sequence comparisons with default parameters. Finally, we checked for best reciprocal hits between genes from each species pair resulting in 451,537 orthologous relationships for all 15 species pairs (see table S2 in the supplementary information).

7.4 Implementation

The algorithms used to generate the transcriptional units were developed in Java and SQL. We used the PostgreSQL 7.4 database (http://www.postgresql.org) for our application. The sets of transcriptional units are available through our web-server (http://bioinfo.amc.uva.nl/HTMseq). Transcriptional View [18] can also be accessed through the web-site (http://bioinfo.amc.uva.nl/human-genetics/transcriptview).

7.5 Results

We constructed and linked transcriptional units for *H. sapiens, M. musculus, R. norvegicus, D. melanogaster, D. rerio, D. simulans, D. yakuba, A. gambiae, B. taurus, C. familiaris, C. elegans, G. gallus, M. domestica, P. troglodytes,* and *T. nigroviridis*. Here we only discuss the results obtained for the human and mouse sets. The full set of transcriptional units is available from our web site and table S1 in the supplementary information provides the overall statistics for the sets for the different organisms.

Using the UCSC genome database builds hg18 and mm8 we derived a set of 33,007 reliable transcriptional units for human and 27,792 transcriptional units for mouse respectively (table 4). The total number of transcriptional units for human and mouse reflects the number of transcriptional units prior to removal of units that map to multiple locations (step 5 of our algorithm). The number of reliable transcriptional units much better reflects the estimated number of human and mouse genes [32]. A large fraction of the transcriptional units contain a RefSeq sequence or a well-defined mRNA. We also observe that only a minor fraction of the transcriptional units contain two or more gene ID's from the NCBI 'Gene' database. Figure 3 shows an example of a transcriptional unit for the human phospholipase A2-activating protein gene (TU number: 315880_s1), which includes two RefSeq sequences and a large number of EST sequences. The mapping of this transcriptional unit on the genomic sequence clearly shows the structure of the gene.

Table 4. Overview of the transcriptional units obtained for *H. sapiens* (hg18) and *M. musculus* (mm8).

		hg18	**mm8**
Total number of TU's		389117	226059
Number of reliable TU	all	33007	27792
	forward strand	16644	13950
	reverse strand	16363	13842
Number of TU's containing RefSeqs		16529	15869
Number of TU's containing mRNA's		22425	15869
Number of TU's containing only EST's		10419	4311
1 exon		5522	2393
2-4 exons		8840	8866
5-9 exons		7124	6891
10-24 exons		8405	7668
>= 25 exons		3116	1974
1 sequence		922	801
2-9 sequences		13452	10624
10-99 sequences		10045	10779
100-999 sequences		7955	5423
1000-9999 sequences		618	163
>= 10000 sequences		15	2
1 gene ID's		18082	20151
2 gene ID's		700	604
3-5 gene ID's		169	90
>= 6 gene ID's		27	26

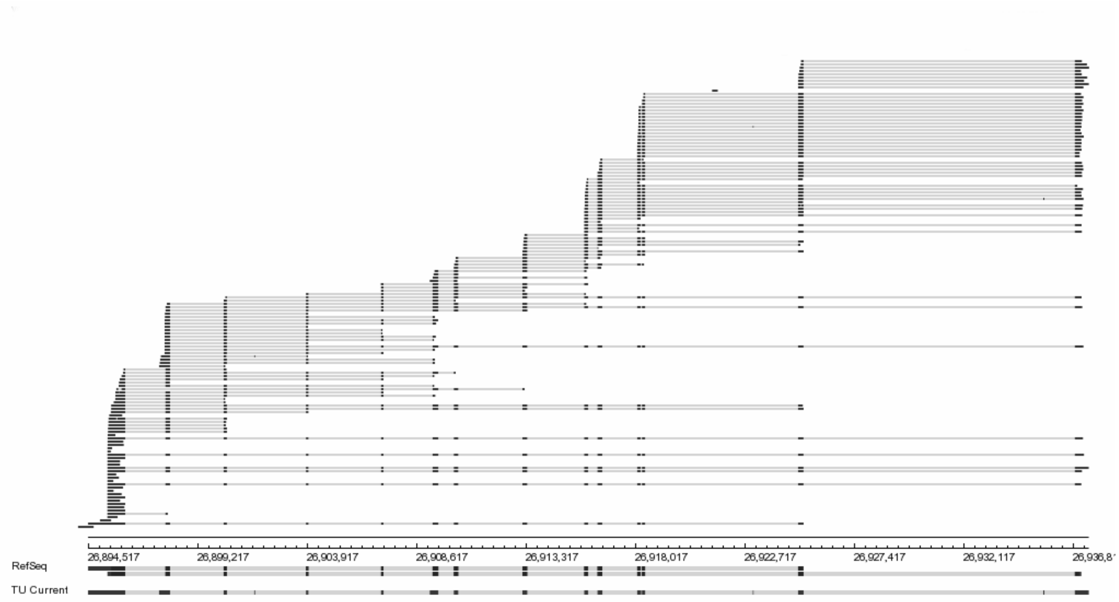


Figure 3. Transcriptional unit for the phospholipase A2-activating protein gene. This transcriptional unit contains two RefSeq sequences and a large number of ESTs. The exon-intron structure is clearly visible.

7.5.1 Cluster size distribution

Table 5 and figure 4 show the distribution of clusters sizes for the human transcriptional units, UniGene, ECgene and AceView. Note that we only included ECgene clusters from the most reliable class A (34,973 clusters; see [10,15] for details). Clearly, we generate the fewest sequence clusters (33,007), which closest approximates the estimated number of 25,947 human genes [32]. The number of clusters generated by UniGene (83,896) largely exceeds the expected number of human genes, which is caused by the large number of singleton clusters that contain only one sequence. Compared with ECgene and AceView, both the Transcriptional Units and UniGene contain more clusters with a (very) large number of sequences.

Table 5. Comparison of human sequence clusters represented by Transcriptional Units (based on hg18), UniGene (build 196), ECgene (based on hg18 build 1, only high confidence clusters) and ACEview (based on human hg17, only high confidence (main) clusters).

Min	Max	TU	UniGene	Ecgene	AceView
32769	65536	1	1	1	0
16385	32768	3	6	2	0
8193	16384	19	22	18	8
4097	8192	48	62	37	44
2049	4096	146	233	136	152
1025	2048	391	739	433	518
513	1024	1024	2141	1395	1762
257	512	2217	4326	3291	3702
129	256	3438	4268	4514	4059
65	128	3430	3376	4084	3104
33	64	2893	3150	3706	2879
17	32	2983	3436	3978	3287
9	15	2514	4061	3610	3476
5	8	3226	5367	4311	5016

3	4	4358	6217	4681	6157
2	2	5394	6068	4899	7263
1	1	922	40423	8673	11427
Total		**33007**	**83896**	**47769**	**52854**

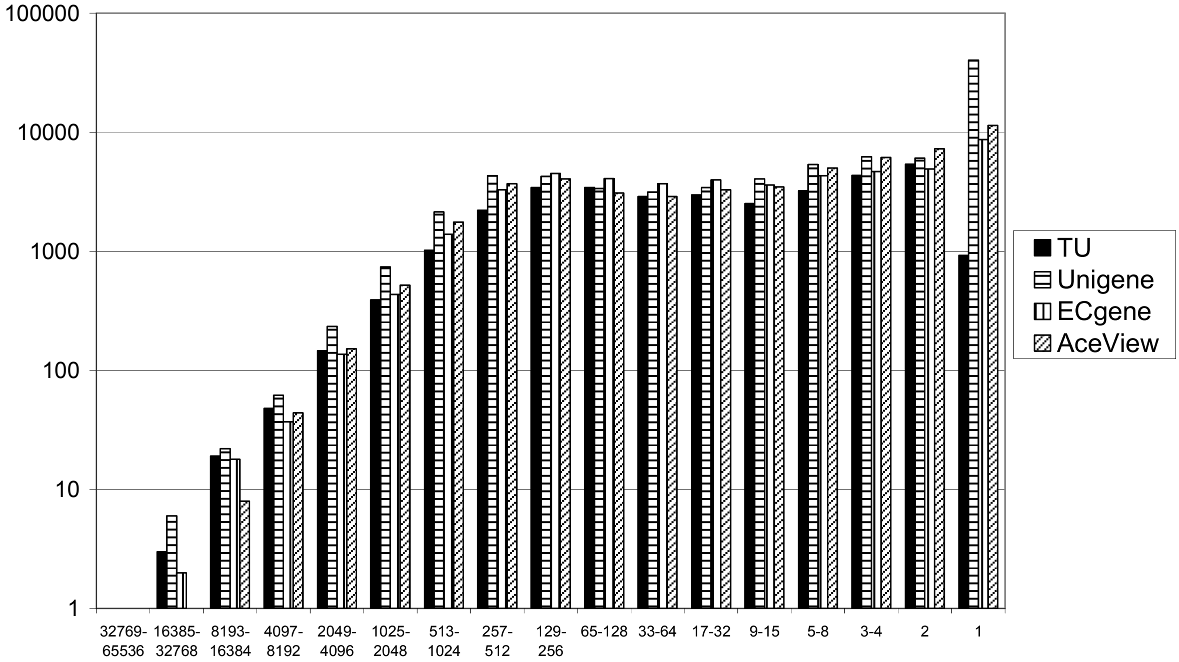


Figure 4. Distribution of cluster sizes for Transcriptional Units and clusters generated by UniGene, ECgene and ACEview. The overall distributions are similar but UniGene generates a large number of singleton clusters.

7.5.2 Hybrid transcriptional units

We also analyzed the number of hybrid transcriptional units, i.e. clusters containing non-overlapping full-length transcripts, and compared these to the number of hybrid clusters obtained with ECgene and AceView. UniGene could not be included in this comparison because the sequence alignments on which UniGene is based and which are required for this analysis were not available.

Table 6. Number of hybrid transcriptional units (based on hg18), ECgene (based on hg18 build 1, only high confidence clusters) and AceView (based on human hg17, only high confidence (main) clusters

Category	**TU**		**ECgene**		**Aceview**	
	# clusters	**%**	**# clusters**	**%**	**# clusters**	**%**
All	33,007	100.00	47,769	100.00	52,935	100.00
>= 1 refseq/mRNA	22,588	68.43	24,168	50.59	26,651	50.35
>= 2 refseq/mRNA	18,440	55.87	19,050	39.88	19,802	37.41
>= 2 refseq/mRNA with > 2.5 Mbp gap	0	0.00	0	0.00	0	0.00
>= 2 refseq/mRNA with > 2.0 Mbp gap	0	0.00	1	0.00	0	0.00
>= 2 refseq/mRNA with > 1.5 Mbp gap	2	0.01	2	0.00	0	0.00
>= 2 refseq/mRNA with > 1.0 Mbp gap	5	0.02	10	0.02	0	0.00
>= 2 refseq/mRNA with > 0.5 Mbp gap	41	0.12	33	0.07	0	0.00

>= 2 refseq/mRNA with > 0.25 Mbp gap	154	0.47	96	0.20	0	0.00
>= 2 refseq/mRNA with > 0.10 Mbp gap	573	1.74	351	0.73	0	0.00
>= 2 refseq/mRNA with > 0.05 Mbp gap	1089	3.30	677	1.42	0	0.00
>= 2 refseq/mRNA with > 0.025 Mbp gap	1759	5.33	1134	2.37	0	0.00
>= 2 refseq/mRNA with > 0.010 Mbp gap	2627	7.96	1706	3.57	1	0.00
>= 2 refseq/mRNA with > 0 Mbp gap	4709	14.27	2676	5.60	299	0.56

Table 6 gives the total number of clusters with at least one full-length transcript and clusters that contain at least two full-length transcripts. For this last category of clusters we determined the amount of separation between these full-length transcripts. The number of hybrid cluster in our set of transcriptional units is roughly similar to the number in ECgene. Our set has fewer hybrid clusters with a very large gap length (> 1 Mbp) whereas ECgene contains fewer hybrid clusters with a gap length smaller than 10 Kbp. AceView contains very few hybrids of any gap length perhaps due to specific manual inspection of these cases.

7.6 Discussion

The development of our algorithms to construct transcriptional units and the comparison of the transcriptional units to clusters obtained with alternative methods demonstrate that this indeed is a very challenging problem. The research presented here made clear that our algorithms do not provide the final solution for the construction of sequence clusters and reveals some of the challenges that still have to be solved in the future. One problem involves the parameterization of the algorithms that affects the final outcome. Some of our choices were made ad hoc based on manual inspection of the result of the clustering. Hence, for many of the (interacting) parameters the effect on the set of transcriptional units has not yet been systematically investigated. This requires one or more objective criteria to judge the overall quality of the sequence clusters but this may be hard to define. Alternatively, one could (semi) manually curate the many clusters obtained for many different organisms. Obviously, this is laborious and inefficient and ultimately also requires objective criteria. During the development of our algorithms we manually inspected many transcriptional units for their structure and validity and changed the (parameterization of the) algorithm to further improve the quality of the sequence clusters. As mentioned in the introduction, we use the set of transcriptional units in many of our applications, which also allows the detection of errors and, subsequently, to improve the algorithms.

The overall procedure to construct transcriptional units is similar to the procedures taken by UniGene, AceView and ECgene but many (details of the) steps in our approach differ from these alternative approaches. Obviously, small changes in the algorithms, lead to differences in outcome with respect to number of clusters, cluster sizes, ability to remove hybrid and pseudo-gene clusters and the number of erroneous clusters in general. For example, Figure 5a shows an example in which our approach was able to generate two separate clusters for FKBP11 and ARF3 genes, while UniGene generated a hybrid cluster containing both genes. Figure 5b shows an example in which UniGene was able to generate two correct clusters for the SF3A2 and AMH genes while our approach resulted in the shown hybrid transcriptional unit. A precise comparison between the different sets of sequence clusters is difficult and would require a detailed comparison of the algorithms, which is not possible due to lack of detailed documentation, the unavailability of the software or sequence alignments.

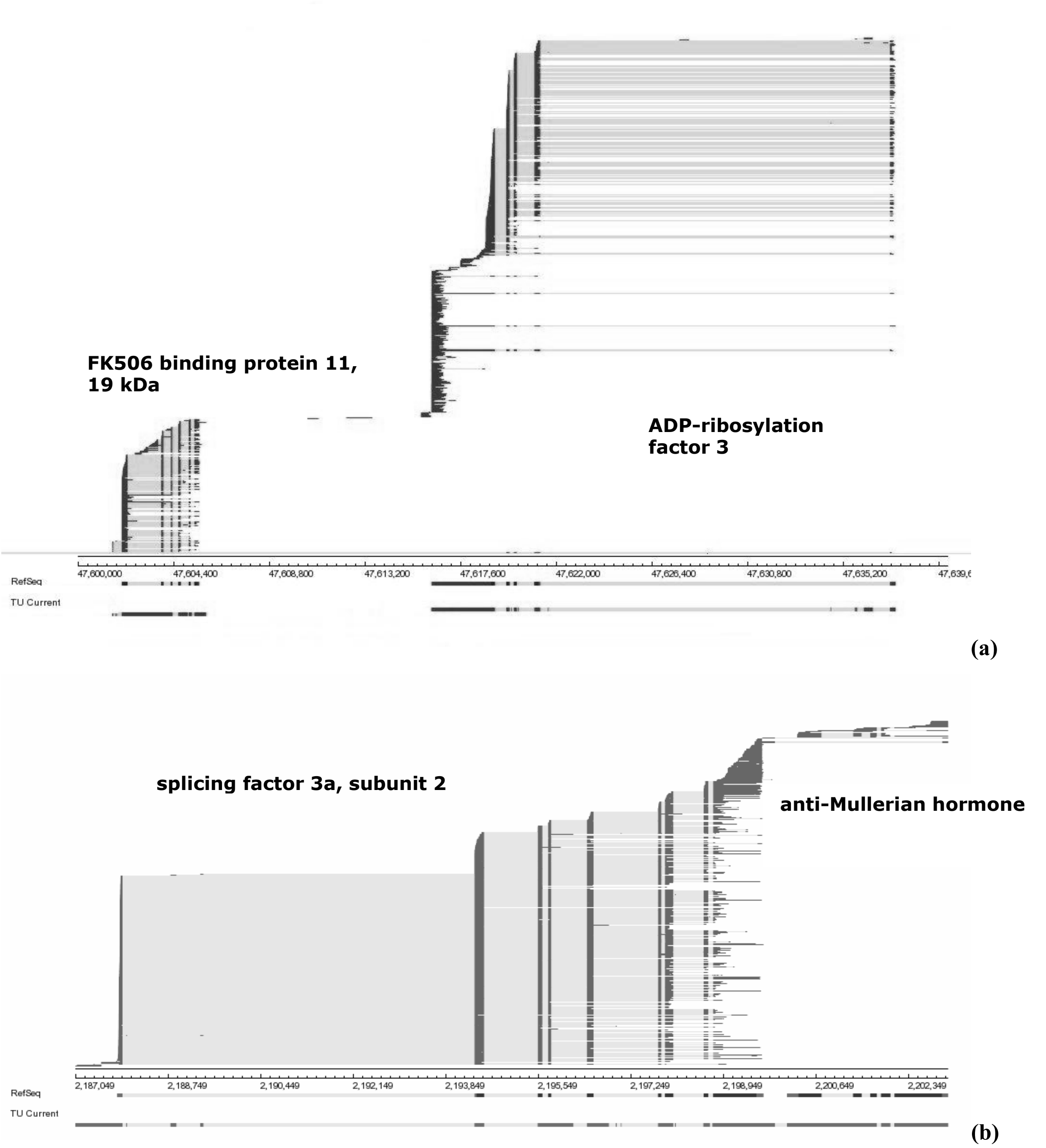


Figure 5. Example hybrid clusters from UniGene and the transcriptional units. **(a)** Two transcriptional units for the FK506 binding protein 11, 19 kDa gene (FKBP11; TU=70186_s1) and ADP-ribosylation factor 3 gene (ARF3; TU=70186_s3). UniGene generates a single erroneous cluster (Hs.119177) that includes both genes. **(b)** One erroneous transcriptional unit (TU number: 144970_s1) that contains both the splicing factor 3a, subunit 2 (SF3A2) and anti-Mullerian hormone (AMH) gene. UniGene correctly generated two separate clusters for these genes (SF3A2 is represented by Hs.115232 & Hs.501353; AMH is represented by Hs.112432).

Another issue concerns the parameterization of the orientation algorithms, which may depend on the organism to which they are applied. There is evidence that different organisms use the canonical polyadenylation signals with different frequencies [21,22-26]. Currently, we optimized the parameterization for human and apply the resulting orientation rules to other organisms assuming that this does not affect the final orientation too much. In addition, from table 3 it was clear that the use of polyadenylation information could not contribute significantly to the orientation of the large number of sequences that did not contain and exon-intron boundary. However, in the ideal situation we would need to derive a set of orientation rules for each organism separately.

It is clear that further work needs to be done to improve sequence clustering algorithms, which seems justified given the many applications in which these clusters are used.

7.7 Supplementary data

The supplementary tables belonging to this article can be found at http://www.cmbi.ru.nl/~timhulse/trscrunits.

7.8 Acknowledgements

This research was supported by the Netherlands organization for scientific research NWO Biomolecular Informatics (NWO-BMI 050.50.201) and NWO Horizon (050-71-404). This work was part of the BioRange programme (project SP1.2.4) of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

7.9 References

1. Galperin MY: The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Res* 2005, 33(Database issue): D5-D24.
2. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J: The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res* 2005, 33 (Database issue): D71-D74.
3. Christoffels A, van Gelder A, Greyling G, Miller R, Hide T, Hide W: STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res* 2001, 29(1): 234-238.
4. Parkinson J, Guiliano DB, Blaxter M: Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics* 2002, 3: 31.
5. Kalyanaraman A, Aluru S, Kothari S, Brendel V: Efficient clustering of large EST data sets on parallel computers. *Nucleic Acids Res* 2003, 31(11): 2963-2974.
6. Malde K, Coward E, Jonassen I: Fast sequence clustering using a suffix array algorithm. *Bioinformatics* 2003, 19(10): 1221-1226.
7. Mudhireddy R, Ercal F, Frank R: Parallel hash-based EST clustering algorithm for gene sequencing. *DNA Cell Biol* 2004, 23(10): 615-623.

8. Frank RL, Ercal F: Evaluation of Glycine max mRNA clusters. *BMC Bioinformatics* 2005, 6(Suppl 2): S7.
9. Ptitsyn A, Hide W: CLU: a new algorithm for EST clustering. *BMC Bioinformatics* 2005, 6(Suppl 2): S3.
10. Kim N, Shin S, Lee S: ECgene: genome-based EST clustering and gene modeling for alternative splicing. *Genome Res* 2005, 15(4): 566-576.
11. Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engstrom PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW et al: Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS.Genet* 2006, 2(4): e62.
12. Thierry-Mieg D, Thierry-Mieg J: AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* 2006, 7(Suppl 1): S12-S14.
13. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: The human genome browser at UCSC. *Genome Res* 2002, 12(6): 996-1006.
14. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ et al: The UCSC Genome Browser Database. *Nucleic Acids Res* 2003, 31(1): 51-54.
15. Kim P, Kim N, Lee Y, Kim B, Shin Y, Lee S: ECgene: genome annotation for alternative splicing. *Nucleic Acids Res* 2005, 33(Database issue): D75-D79.
16. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA et al: The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 2001, 291(5507): 1289-1292.
17. Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH: The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* 2003, 13(9): 1998-2004.
18. Valentijn LJ, Koster J, Versteeg R: Read-through transcript from NM23-H1 into the neighboring NM23-H2 gene encodes a novel protein, NM23-LV. *Genomics* 2006, 87(4): 483-489.
19. Burset M, Seledtsov IA, Solovyev, VV: Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* 2000, 28(21): 4364-4375.
20. Wickens M: How the messenger got its tail: addition of poly(A) in the nucleus. *Trends Biochem. Sci* 1990, 15(7): 277-281.
21. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D: Patterns of variant polyadenylation signal usage in human genes. *Genome Res* 2000, 10(7): 1001-1010.
22. Sheets MD, Ogg SC, Wickens MP: Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res* 1990, 18(19): 5799-805.
23. Gautheret D, Poirot O, Lopez F, Audic S, Claverie JM: Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res* 1998, 8(5): 524-530.
24. Graber JH, Cantor CR, Mohr SC, Smith TF: In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc.Natl.Acad.Sci.U.S.A* 1999, 96 (24): 14055-14060.

25. Beaudoing E, Gautheret D: Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res* 2001, 11(9): 1520-1526.
26. van Kampen AH, Waaijer R et al: The identification of Serial Analysis of Gene Expression (SAGE) tags. *Manuscript in preparation*.
27. Tabaska JE, Zhang MQ: Detection of polyadenylation signals in human DNA sequences. *Gene* 1999, 231(1-2): 77-86.
28. Yu P, Ma D, Xu M: Nested genes in the human genome. *Genomics* 2005, 86(4): 414-422.
29. Sorek R, Safer HM: A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res* 2003, 31(3): 1067-1074.
30. Maglott D, Ostell J, Pruitt KD, Tatusova T: Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005, 33(Database issue): D54-D58.
31. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006, 7(4): R31.
32. Pennisi E: Human genome. A low number wins the GeneSweep Pool. *Science* 2003, 300(5625): 1484.

Chapter 8

General discussion

The introduction of this thesis gives an impression of the growing importance of genomics in general and orthology in particular, and shows how the field of genomics can be connected to the fields of drug discovery and toxicology. The most promising aspect of this connection lies in the genomics-based discovery of biomarkers: characteristics that are objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention [1]. However, the application of genomics in drug discovery is not an easy one, and has still not yielded the great results it was expected to give a few years ago. In this thesis we try to help solving the problems encountered when applying genomics methods in drug discovery. The first chapters contain fundamental studies concerning orthology, sequence comparisons and phylogenetic patterns, designed to form a guideline for the application of these methods. The other chapters show the application of orthology in several fields: immunology, evolutionary biology and transcriptomics. In this discussion the conclusions from all these chapters will be dealt with, and we will discuss the upcoming field of pharmacophylogenomics [2], the intersection of pharmaceutics and phylogenomics: can it really shorten the drug discovery pipeline?

Before we can use genomics and orthology in the drug discovery pipeline, it is important to look at their fundamental aspects. Concerning orthology, numerous methods for ortholog identification exist, and we chose that our first step should be to test some of these methods on how well they perform in terms of functional similarity [3]. The assumption that orthologs have a highly similar function is important because it forms the basis for the main goal of ortholog identification: the transfer of functional annotation of proteins from one species to the other. However, numerous cases have been described where orthologs have, in fact, different functions. The correct definition of orthology is ‘the evolutionary relationship between homologous genes whose independent evolution reflects a speciation event’ [4]. Usually these proteins maintain a similar function, but this is not always the case. This discrepancy between the official definition of orthology and the way it is used (functional equivalence), can give problems within the field of pharmacophylogenomics, because we are mainly interested in the functional equivalence and not directly in the evolutionary origin of the studied genes or proteins. This is reflected in the fact that only some of the benchmarked ortholog identification methods really use evolutionary methods such as phylogenetic trees [5], whereas most methods just use sequence comparisons [6, 7]. Moreover, the quality of an ortholog determination method depends not only on the method itself, but also on its settings and other programs used in the process. Using for example either the heuristic method BLAST [8] or the Smith-Waterman implementation of ParAlign [9] as the sequence comparison program within the ortholog identification process, can give quite different results [10]. The same holds for the multiple sequence alignment algorithm that is used if a ortholog identification method requires multiple alignments and phylogenetic trees: ClustalW [11], MUSCLE [12], or any other algorithm can give very different multiple alignments, resulting in different phylogenetic trees and different orthology assignments. One of the applications of orthologous relationships lies in the creation of orthologous groups and the subsequent creation of phylogenetic patterns [13], which display the presence or absence of certain genes over a set of species [14]. These patterns can be used, for example, to cluster genes that occur in the same species or taxons. These genes are likely to have a similar function or to be involved in the same biological process. Phylogenetic patterns can also be used to study gene families and their expansions or deletions over time, which can be very useful in drug discovery because interspecies differences in drug response could be explained by looking at expansions or deletions of certain genes in the pathway.

The combination of genomics and pharmacology is usually referred to as ‘pharmacogenomics’. As defined a few years ago [15]: “pharmacogenomics refers to the general study of all of the many different genes that determine drug behavior”. However, the term is sometimes mixed up with ‘pharmacogenetics’: the study of inherited differences (variation) in drug metabolism and response. The usage of orthology to find solutions for drug discovery problems is part of pharmacogenomics, but is not part of pharmacogenetics. Orthology can only provide answers for interspecies differences, not interhuman differences. The main application of orthology in drug discovery lies in the assumption that differences in drug response between human and model organisms can be explained by looking at the orthologous proteins between these species. In this thesis we tried to show that single proteins, and even protein families or complete protein pathways, can be linked cross-species by identifying orthologous relationships. The study of the evolution of the immune system from model organisms (chicken, rat, mouse, etc.) to man, described in chapter 5, is a good example of this. For drug discovery the interspecies mapping of protein pathways is of specific interest. A different response to a certain drug in man and in a model organism can be elucidated by mapping the organisms’ pathways onto each other. This could increase the predictive value of studies in animal models drastically. However, studies like this need highly accurate and reliable orthology information. Moreover, other factors like alternative transcripts, expression levels and three-dimensional structure could be part of the solution. An example of this is the (unpublished) study we did on the thrombin/trypsin inhibition pathway [16]. If thrombin inhibitors are administred to rats, side effects occur because these inhibitors also act on the trypsin inhibition pathway. The cholecystokinine (CCK) levels in the rats rise, which overstimulates the pancreas, leading to pancreatic tumors. In mouse these CCK levels rise much less, and in man these levels do not rise at all. This makes testing of drugs related to this pathway, such as Exanta/ximelagatran [17], rather difficult. We tried to explain the interspecies differences by looking at the (numbers of) orthologs in the trypsin inhibition pathway. Although we found some differences in the pathway over these three species, these could not be the only reason for the differences in CCK response. A next step should be the use of expression data and structural data, an approach that has been useful in recent studies [18-20]. All in all, in order to provide an answer to pharmacogenomics questions, a whole range of genomics data might be needed, instead of just orthology data.

Orthology has a wide range of applications: immunology studies [21], evolutionary studies [22] and transcriptomics studies [23] all benefit from the use of orthologous relationships. However, application of orthology alone is not likely to answer many research questions. As is shown in especially the immune system and trypsin inhibition studies, a wide range of functional genomics data needs to be gathered to shed light on complex systems such as pathways in the field of drug discovery. This makes the field of pharmacophylogenomics a difficult one, but with the ever growing availability of genomics data it will certainly have its benefits on the long term. This especially holds for the availability of more genomics data from existing model organisms such as dog, macaque and goat. In the future, the completeness and higher reliability of genomics data will enable researchers to perform studies like in this thesis in more detail and with more accuracy. The best future results will be obtained when the knowledge acquired from this thesis is combined with expression data and structural data. Finally, findings should be directly incorporated in clinical studies, as seen in translational medicine [24]. This will fasten the drug discovery pipeline significantly.

References

1. Frank R, Hargreaves R: Clinical biomarkers in drug discovery and development. *Nat Rev Drug Discov* 2003, 2(7):566-580.
2. Searls DB: Pharmacophylogenomics: genes, evolution and drug targets. *Nat Rev Drug Discov* 2003, 2(8):613-623.
3. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006, 7(4):R31.
4. Fitch WM: Distinguishing homologous from analogous proteins. *Syst Zool* 1970, 19(2):99-113.
5. van Noort V, Snel B, Huynen MA: Predicting gene function by conserved co-expression. *Trends Genet* 2003, 19(5):238-242.
6. Li L, Stoeckert CJ, Jr., Roos DS: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003, 13(9):2178-2189.
7. Remm M, Storm CE, Sonnhammer EL: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001, 314(5):1041-1052.
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215(3):403-410.
9. Rognes T: ParAlign: a parallel sequence alignment algorithm for rapid and sensitive database searches. *Nucleic Acids Res* 2001, 29(7):1647-1652.
10. Hulsen T, de Vlieg J, Leunissen JA, Groenen PM: Testing statistical significance scores of sequence comparison methods with structure similarity. *BMC Bioinformatics* 2006, 7:444.
11. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994, 22(22):4673-4680.
12. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004, 32(5):1792-1797.
13. Hulsen T, de Vlieg J, Groenen PM: PhyloPat: phylogenetic pattern analysis of eukaryotic genes. *BMC Bioinformatics* 2006, 7:398.
14. Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000, 28(1):33-36.
15. Mehr IJ: Preparing for the revolution--pharmacogenomics and the clinical lab. *Pharmacogenomics* 2000, 1(1):1-4.
16. Hulsen T: Thrombin/trypsin inhibition pathway study; unpublished data.
17. Hopfner R: Ximelagatran (AstraZeneca). *Curr Opin Investig Drugs* 2002, 3(2):246-251.
18. Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RH: Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* 2006, 38(12):1375-1377.
19. Stephenson K: Sec-dependent protein translocation across biological membranes: evolutionary conservation of an essential protein transport pathway (review). *Mol Membr Biol* 2005, 22(1-2):17-28.

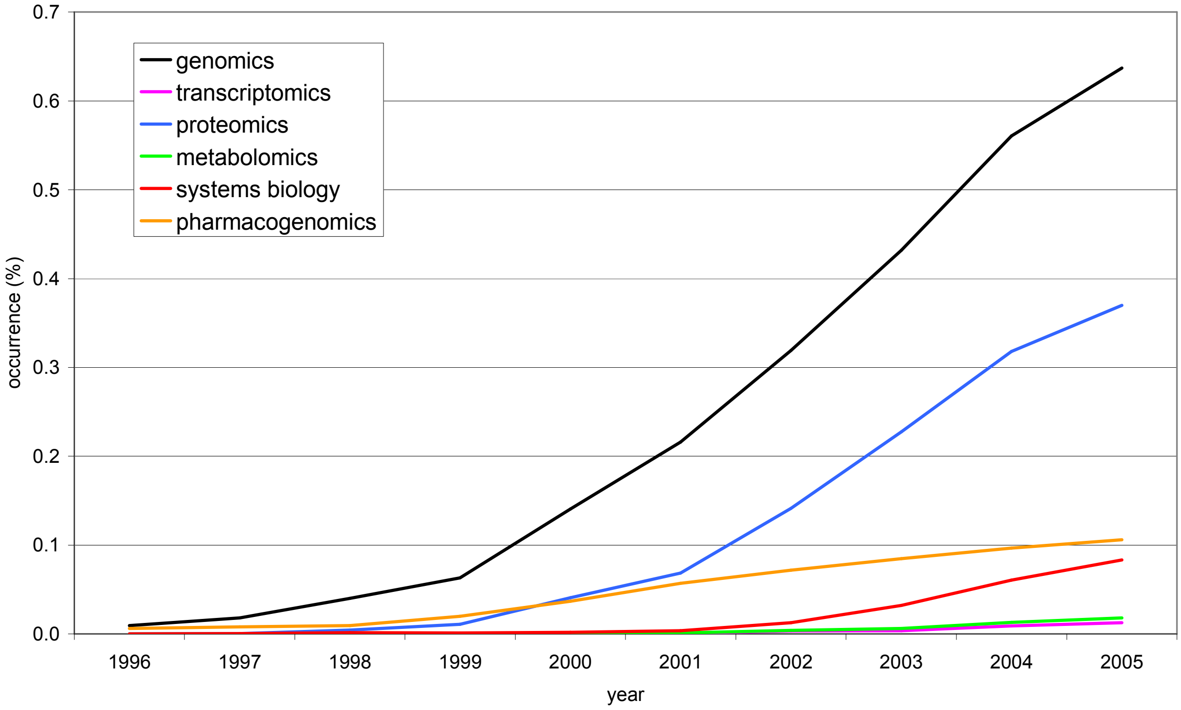
20. Mao F, Su Z, Olman V, Dam P, Liu Z, Xu Y: Mapping of orthologous genes in the context of biological pathways: An application of integer programming. *Proc Natl Acad Sci U S A* 2006, 103(1):129-134.
21. Hulsen T, Fleuren WWM, Kerstens HHD, Groenen MAM, Groenen PMA: Evolution of the immune system from model organism to man. *Manuscript in preparation*.
22. Franck E, Hulsen T, Huynen MA, Lubsen NH, de Jong WW, Madsen O: Dynamics of head-to-head genes in vertebrates. *Manuscript in preparation*.
23. van Hooff S, Monajemi R, Hulsen T, van Batenburg MF, Versteeg R, van Kampen AHC: Transcriptional Units: a set of gene-oriented sequence clusters. *Submitted*.
24. Fitzgerald GA: Opinion: anticipating change in drug development: the emerging era of translational medicine and therapeutics. *Nat Rev Drug Discov* 2005, 4(10):815-818.

List of abbreviations

Ag	*Anopheles gambiae*
Am	*Apis mellifera*
AP	Average Precision
BBH	Best Bidirectional Hit
bf z	Biofacet with Z-score
BLAST	Basic Local Alignment Tool
bl e	BLAST with e-value
BRH	Best Reciprocal Hit
BLOSUM	Blocks Substitution Matrix
Bt	*Bos taurus*
Ce	*Caenorhabditis elegans*
Cf	*Canis familiaris*
Ci	*Ciona intestinalis*
CluSTr	Clusters of SWISS-PROT and TrEMBL
CMBI	Center for Molecular and Biomolecular Informatics
COG	Clusters of Orthologous Groups
CVE	Coverage Versus Error
DIP	Database of Interacting Proteins
Dm	*Drosophila melanogaster*
Dr	*Danio rerio*
EMBL	European Molecular Biology Laboratory
EPPS	Extended Phylogenetic Pattern Search
EST	Expressed Sequence Tag
fa e	FASTA with e-value
FFP	First False Positive
Gg	*Gallus gallus*
GO	Gene Ontology
HGNC	HUGO Gene Nomenclature Committee
Hs	*Homo sapiens*
HTM	Human Transcriptome Map
HTML	HyperText Markup Language
HUGO	Human Genome Organisation
INP	InParanoid
INPB	InParanoid, Best scoring pair
KOG	euKaryotic Orthologous Groups
KOGB	euKaryotic Orthologous Groups, Best scoring pair
MBRH	Multiple Best Reciprocal Hit
MCL	Markov Cluster Algorithm

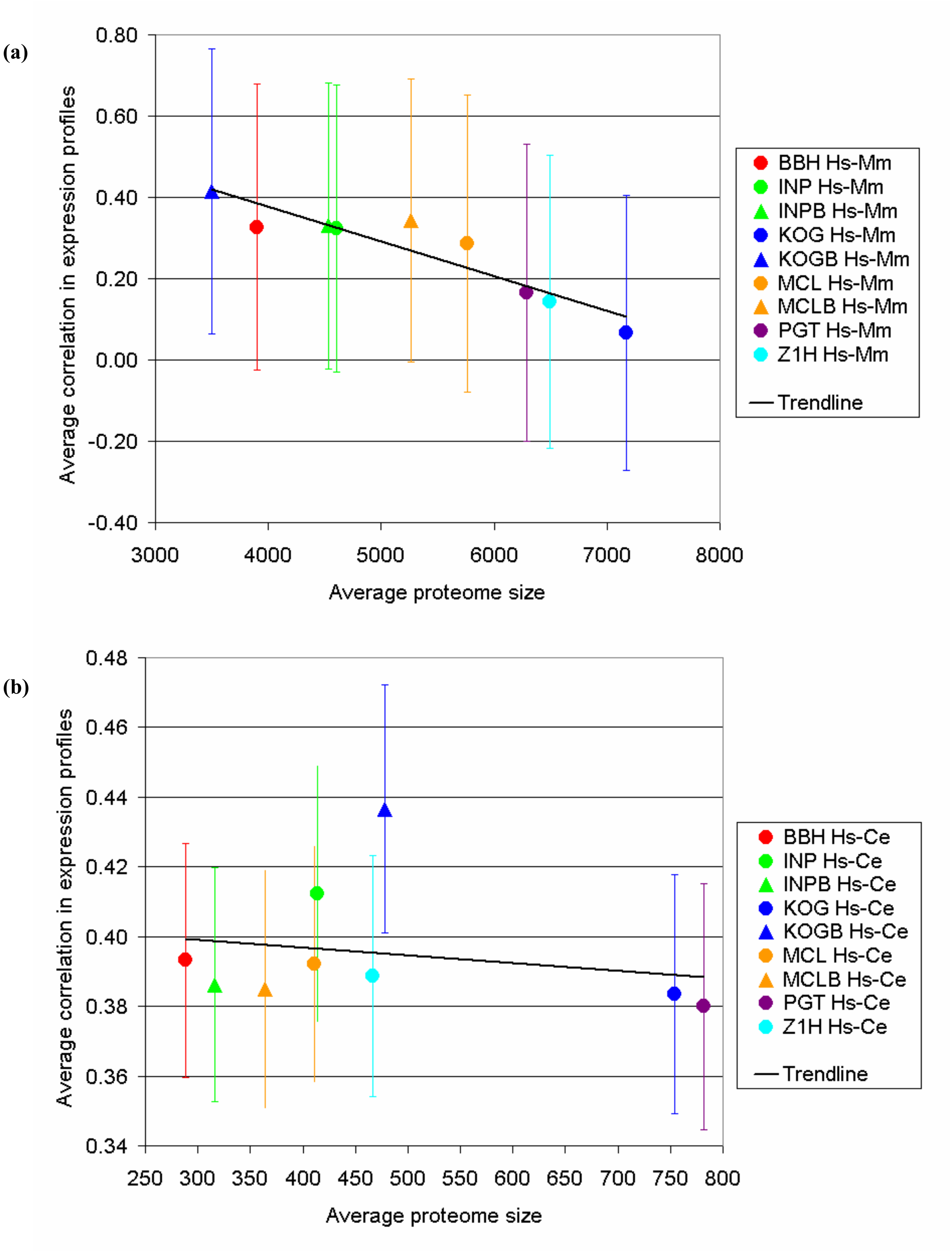
MCLB	Markov Cluster Algorithm, Best scoring pair
Md	*Monodelphis domestica*
Mm	*Mus musculus*
Mmul	*Macaca mulatta*
Mmus	*Mus musculus*
mRNA	Messenger RiboNucleic Acid
MTM	Mouse Transcriptome Map
Mu	*Macaca mulatta*
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
MySQL	My Structured Query Language
NCBI	National Center for Biotechnology Information
NTP	Number of True Positives
pa e	ParAlign with e-value
pc e	Paracel with e-value
PGT	PhyloGenetic Tree
PhIG	Phylogenetically Inferred Group
PHYLIP	Phylogeny Inference Package
PPS	Phylogenetic Pattern Search
Pt	*Pan troglodytes*
RHS	Reciprocal Hit based on Synteny information
Rn	*Rattus norvegicus*
ROC	Receiver Operating Characteristic
Sc	*Saccharomyces cerevisiae*
SCOP	Structural Classification of Proteins
SNOMED	Systematized Nomenclature Of Medicine
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
ss e	SSEARCH with e-value
SW	Smith-Waterman
Tn	*Tetraodon nigroviridis*
Tr	*Takifugu rubripes*
UBRH	Unique Best Reciprocal Hit
Xt	*Xenopus tropicalis*
Z1H	Z 1 Hundred

Color figures



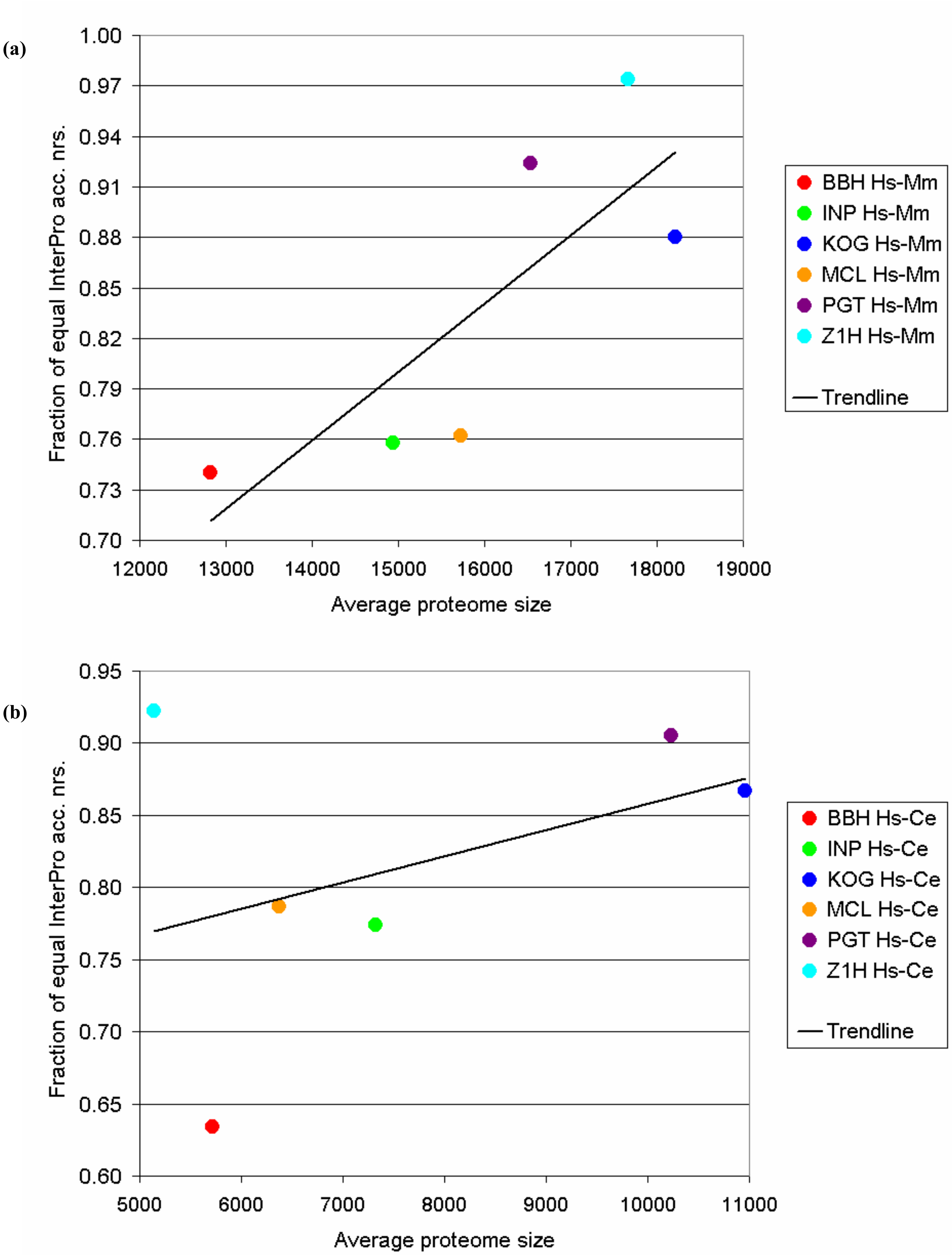
Chapter 1, figure 1. Popularity of -omics search terms in the PubMed database

The percentage of articles (titles + abstracts) in the PubMed database that contain the words 'genomics' (black line), 'transcriptomics' (purple line), 'proteomics' (blue line), 'metabolomics' (green line), 'systems biology' (red line) or 'pharmacogenomics' (orange line). Horizontal axis: year. Vertical axis: number of articles that contain that specific search term, divided by the total number of articles published in that year (in %).



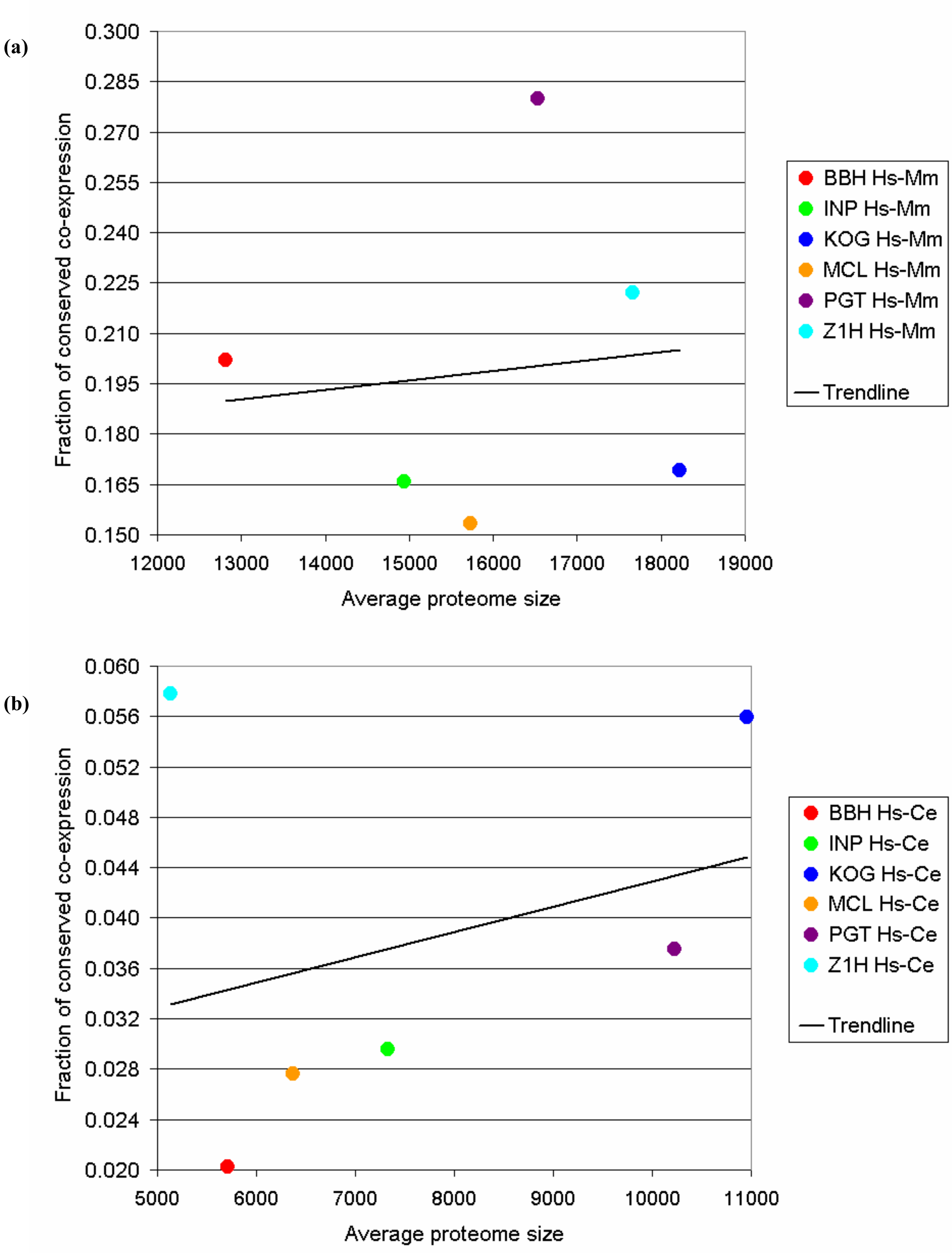
Chapter 2, figure 1. Correlation in expression profiles

Correlation in expression patterns between the **(a)** human-mouse (Hs-Mm) and **(b)** human-worm (Hs-Ce) orthologous pairs from the benchmarked methods versus the average proteome size. Vertical error bars show the standard deviation from the average correlation coefficient. The trendline shown is a linear regression trendline. The methods having a fourth letter 'B' behind the method name, shown as squares in the graph, are group orthology methods in which only the best scoring pairs are taken into account. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*.



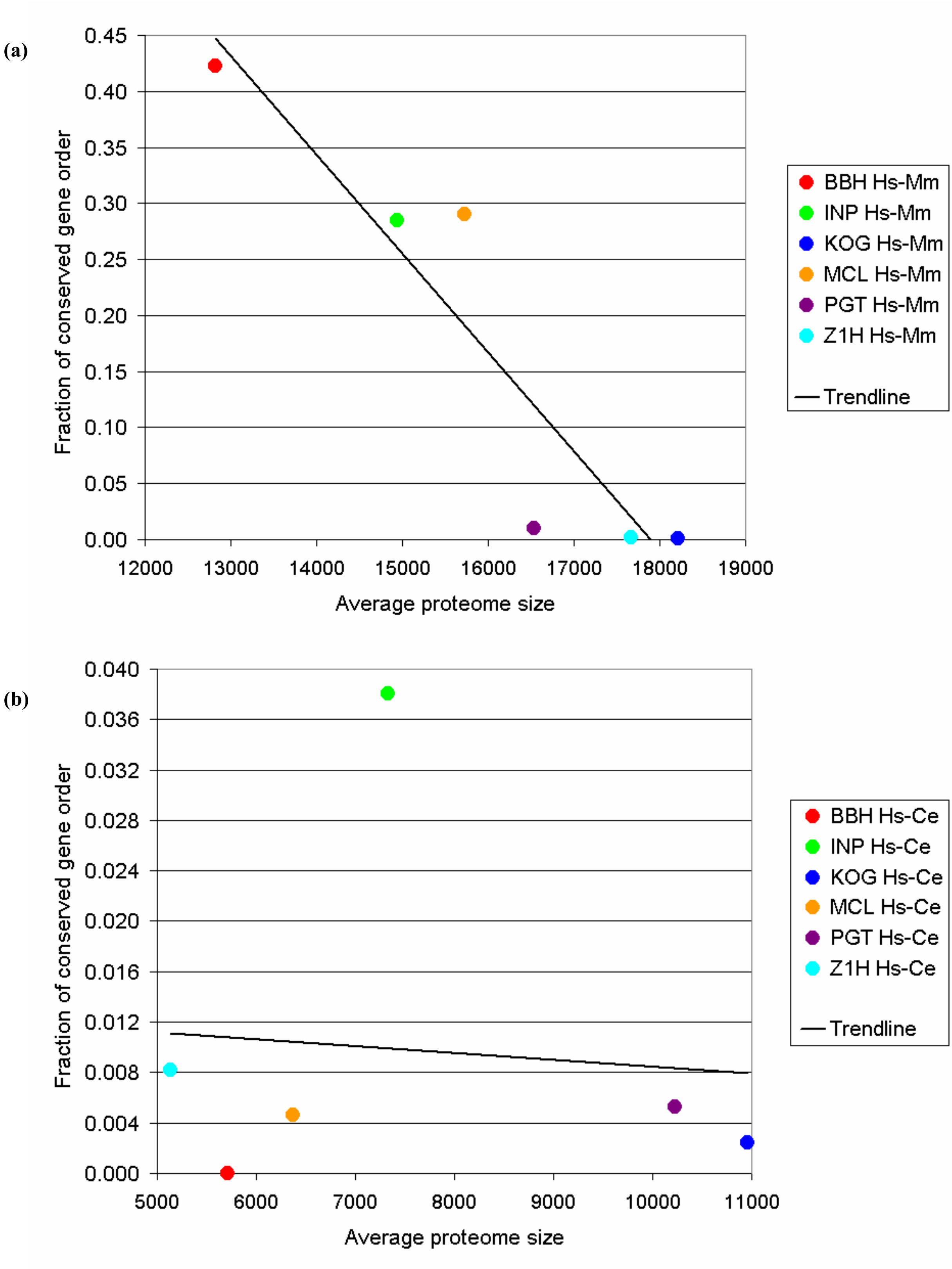
Chapter 2, figure 2. Equal InterPro accession number

Conservation of InterPro accession number between the **(a)** human-mouse (Hs-Mm) and **(b)** human-worm (Hs-Ce) orthologous pairs from the benchmarked methods versus the average proteome size. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*.



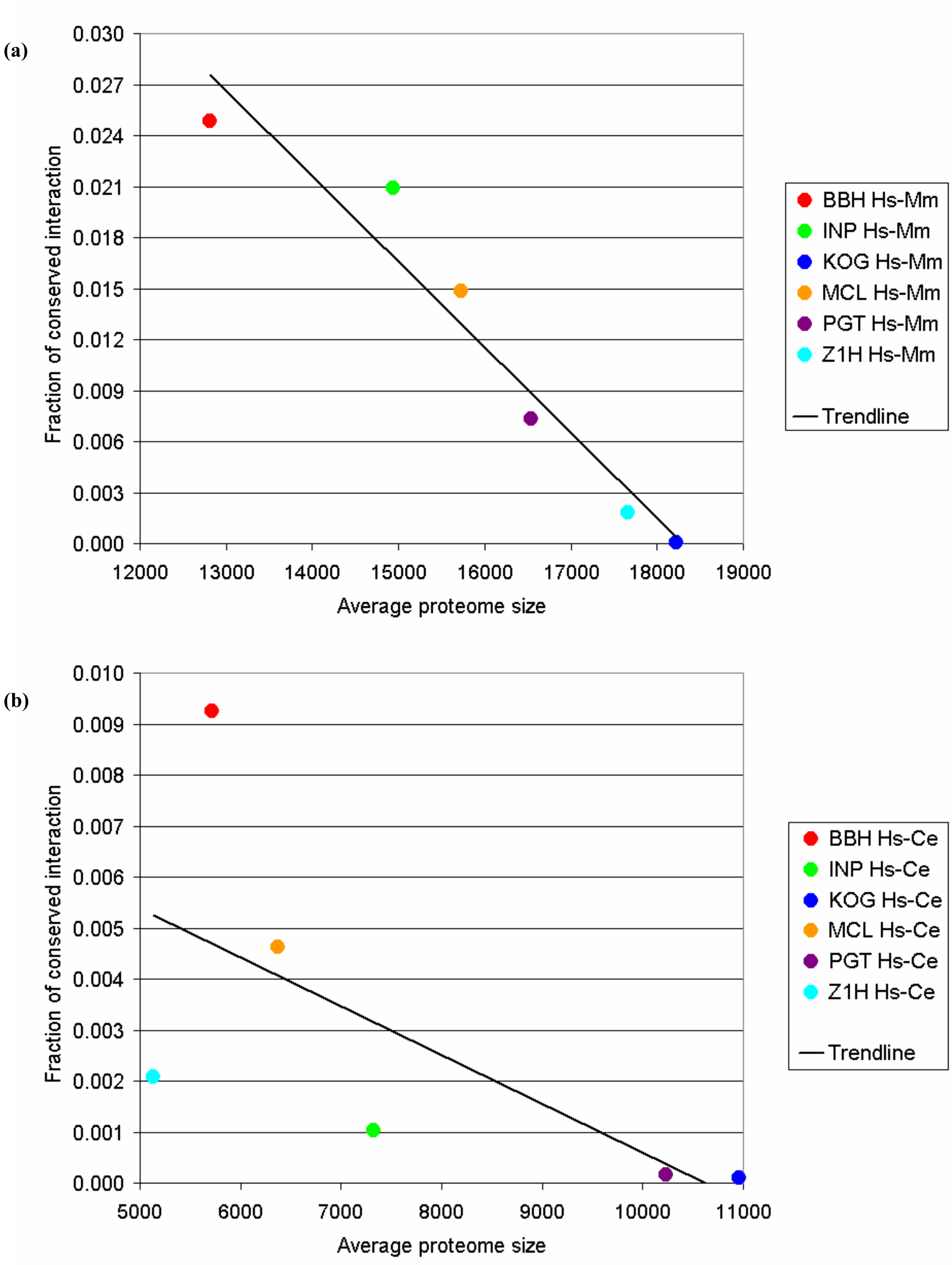
Chapter 2, figure 3. Conservation of co-expression

Conservation of co-expression from human-human gene pairs to orthologous **(a)** mouse-mouse and **(b)** worm-worm gene pairs from the benchmarked methods versus the average proteome size. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*.



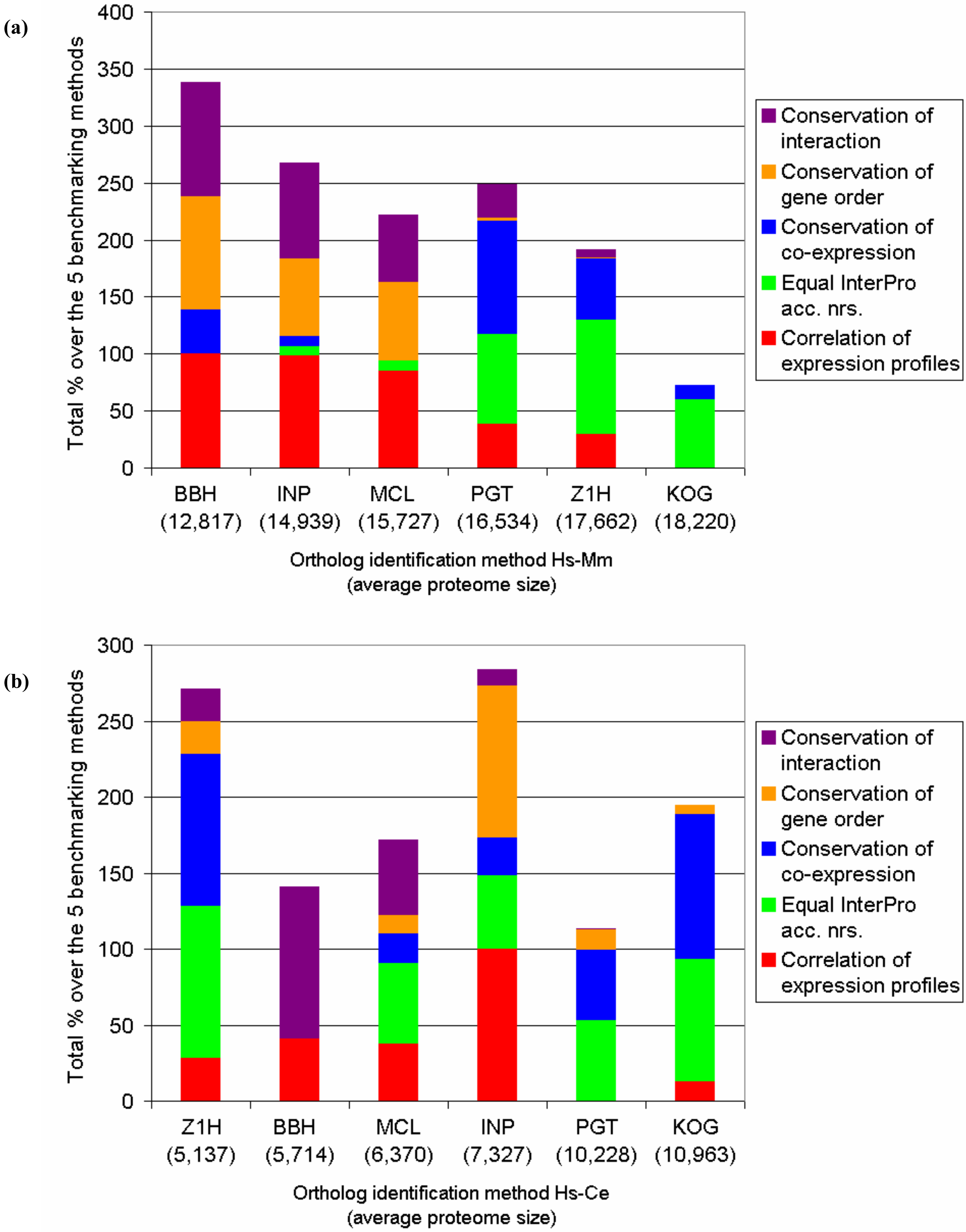
Chapter 2, figure 4. Conservation of gene order

Conservation of gene order from human-human gene pairs to orthologous **(a)** mouse-mouse and **(b)** worm-worm gene pairs from the benchmarked methods versus the average proteome size. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*.



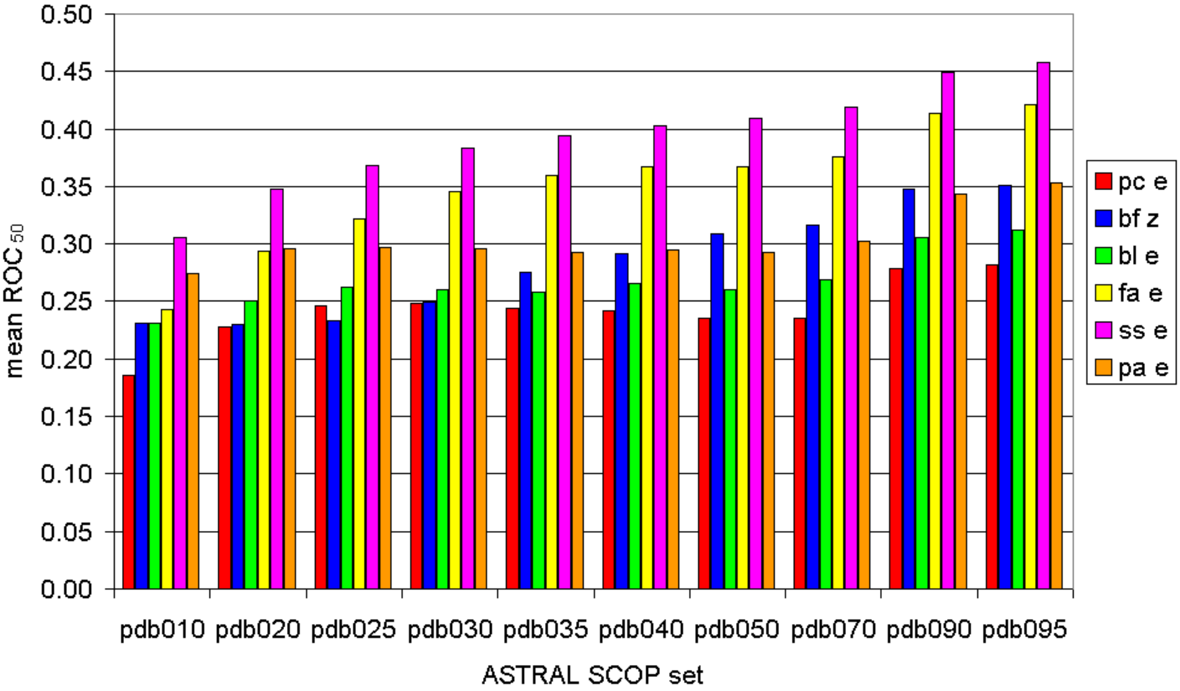
Chapter 2, figure 5. Conservation of protein-protein interaction

Conservation of protein-protein interaction from human-human protein pairs to orthologous **(a)** mouse-mouse and **(b)** worm-worm protein pairs from the benchmarked methods versus the average proteome size. Ce, *Caenorhabditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*.



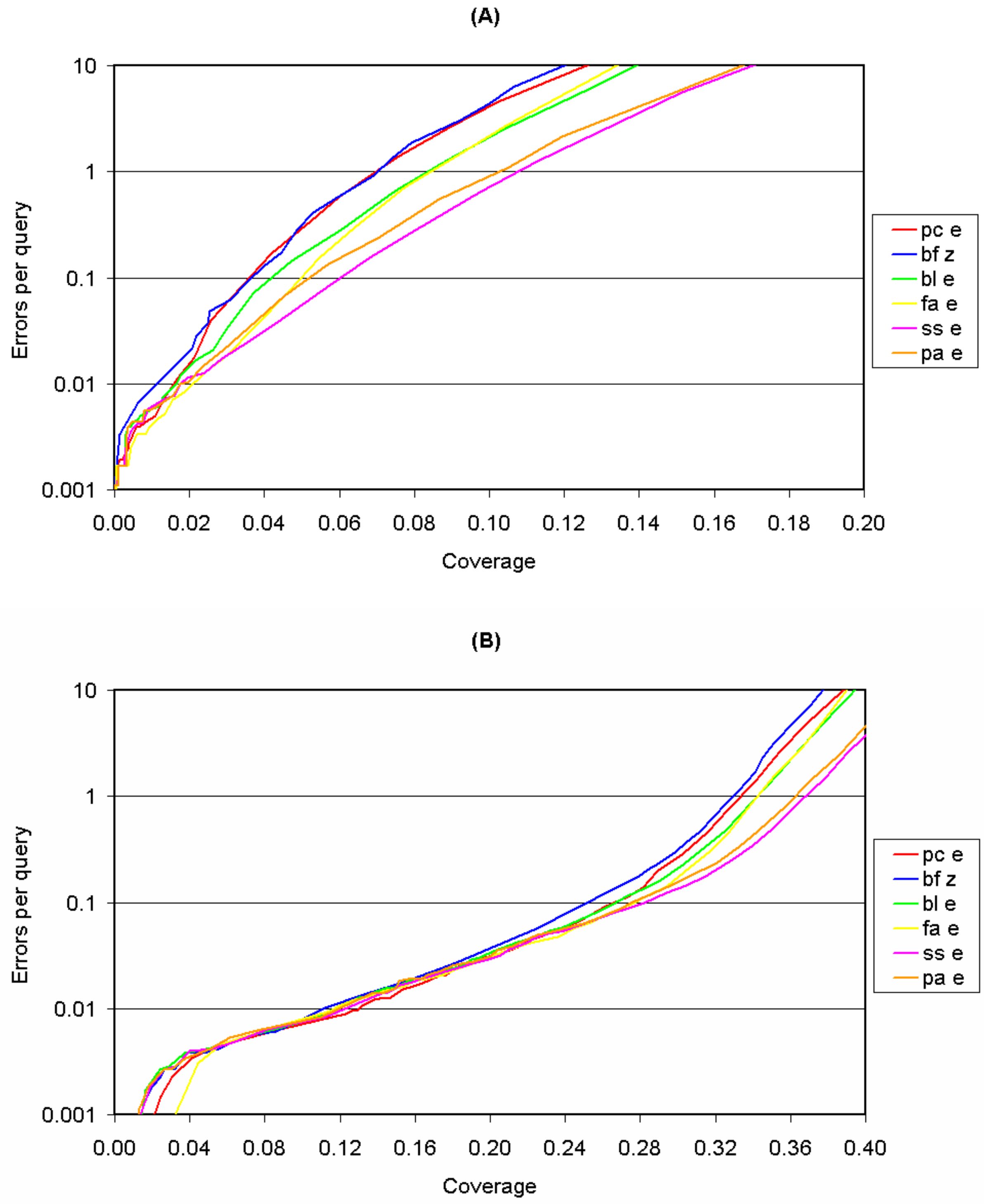
Chapter 2, figure 6. Overall scoring graph

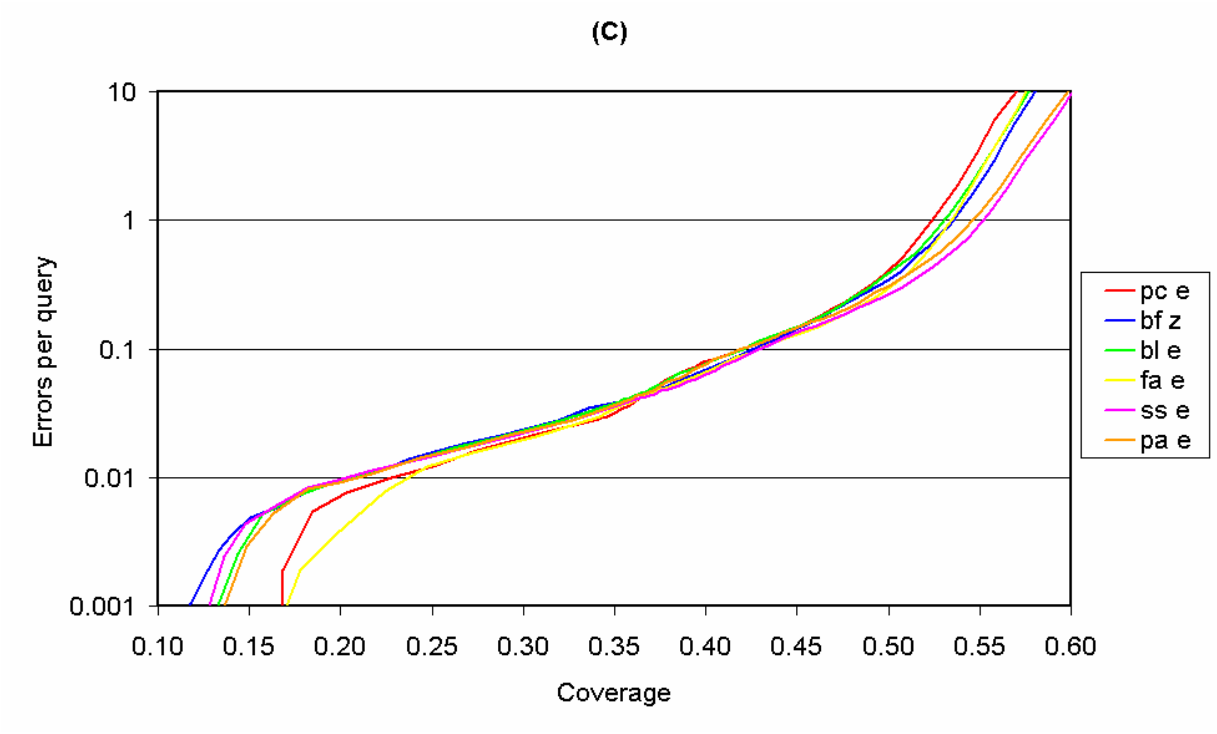
Overall scoring graph, created by adding up all normalized benchmarking scores per ortholog identification method. X-axis, the several ortholog identification methods, sorted by average proteome size or number of protein pairs; Y-axis, the sum of all five benchmarking scores per ortholog identification method. Red, correlation of expression profiles; green, equal InterPro accession numbers; blue, conservation of co-expression; orange, conservation of gene order; purple, conservation of protein-protein interaction. **(a)** Human-mouse (Hs-Mm). **(b)** Human-worm (Hs-Ce).



Chapter 3, figure 1. The mean Receiver Operating Characteristic scores for ten different ASTRAL SCOP sets

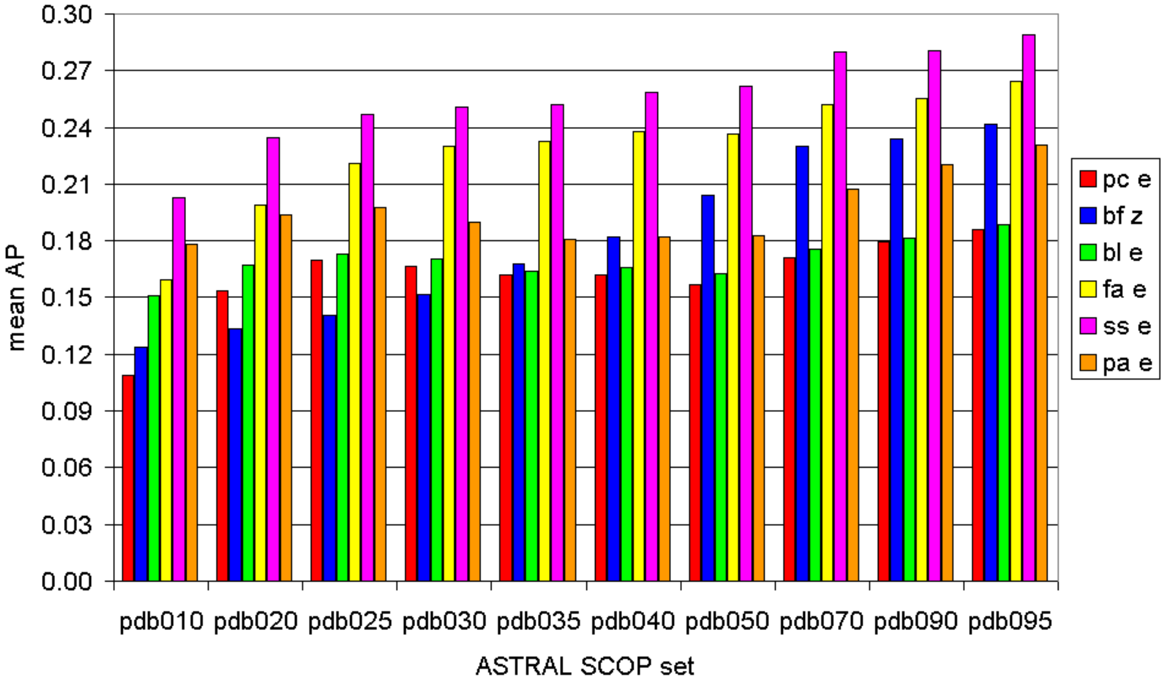
The maximal structural identity percentage of each set increases from the left to the right, from 10% to 95%. Red bars: mean ROC$_{50}$ scores calculated using the Paracel Smith-Waterman algorithm. Blue bars: mean ROC$_{50}$ scores calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Green bars: mean ROC$_{50}$ scores calculated using the BLAST algorithm. Yellow bars: mean ROC$_{50}$ scores calculated using the FASTA algorithm. Purple bars: mean ROC$_{50}$ scores calculated using the SSEARCH algorithm. Orange bars: mean ROC$_{50}$ scores calculated using the ParAlign Smith-Waterman algorithm.


(A)
Errors per query
10
1
0.1
0.01
0.001
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
Coverage
pc e
bf z
bl e
fa e
ss e
pa e
(B)
Errors per query
10
1
0.1
0.01
0.001
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.32
0.36
0.40
Coverage
pc e
bf z
bl e
fa e
ss e
pa e



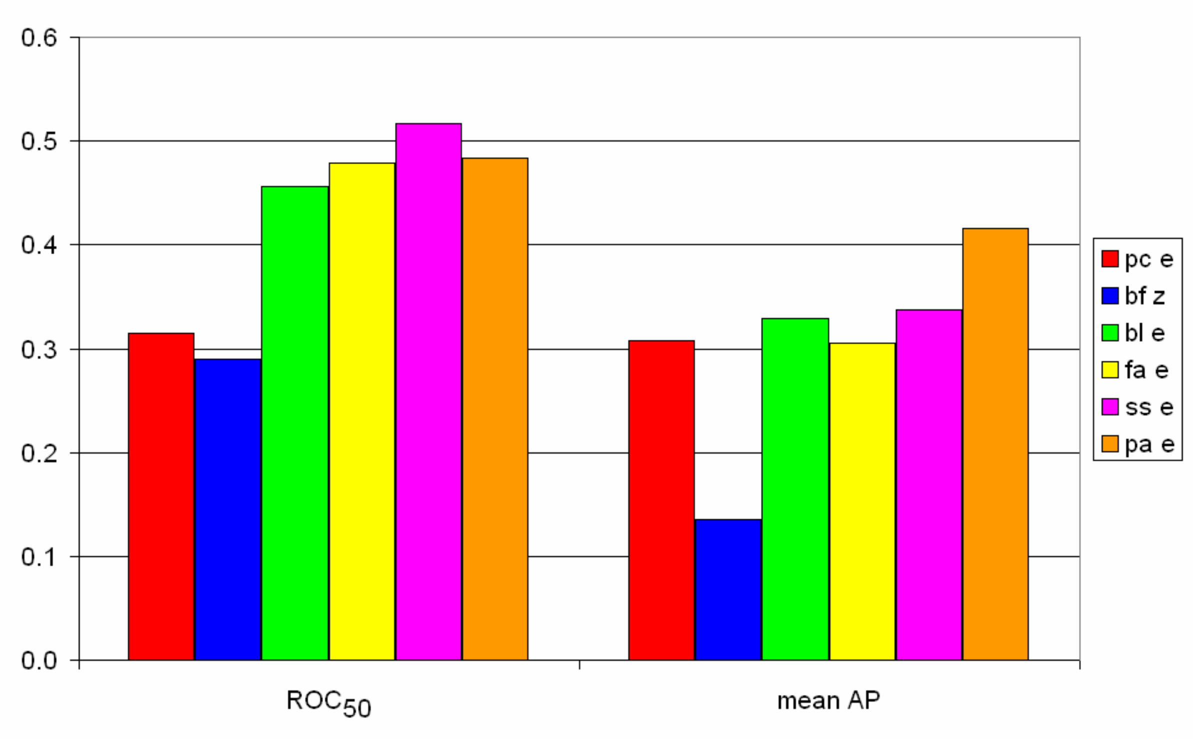
Chapter 3, figure 2. (a) Coverage versus error plot for the ASTRAL SCOP PDB010 set. **(b)** Coverage versus error plot for the ASTRAL SCOP PDB035 set. **(c)** Coverage versus error plot for the ASTRAL SCOP PDB095 set.

Red line: calculated using the Paracel Smith-Waterman algorithm. Blue line: calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Green line: calculated using the BLAST algorithm. Yellow line: calculated using the FASTA algorithm. Purple line: calculated using the SSEARCH algorithm. Orange line: calculated using the ParAlign Smith-Waterman algorithm.



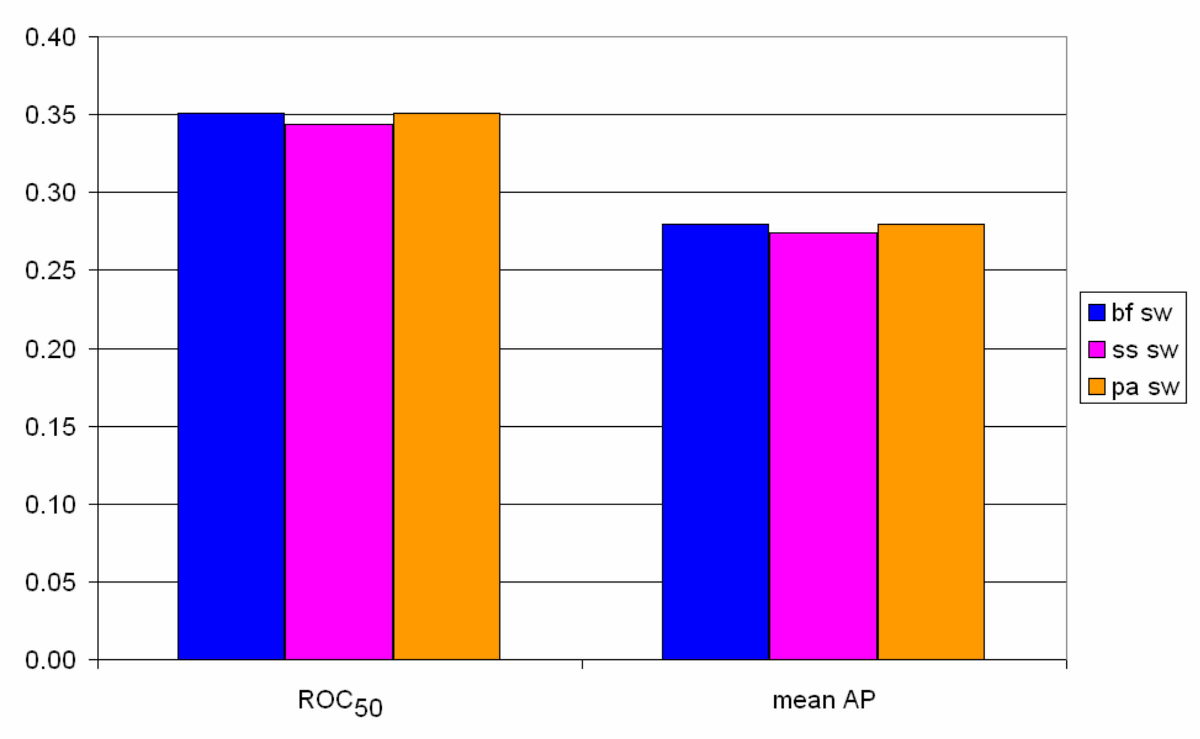
Chapter 3, figure 3. The average precision values for ten different ASTRAL SCOP sets

The maximal structural identity percentage of each set increases from the left to the right, from 10% to 95%. Red bars: mean AP values calculated using the Paracel Smith-Waterman algorithm. Blue bars: mean AP values calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Green bars: mean AP values calculated using the BLAST algorithm. Yellow bars: mean AP values calculated using the FASTA algorithm. Purple bars: mean AP values calculated using the SSEARCH algorithm. Orange bars: mean AP values calculated using the ParAlign Smith-Waterman algorithm.



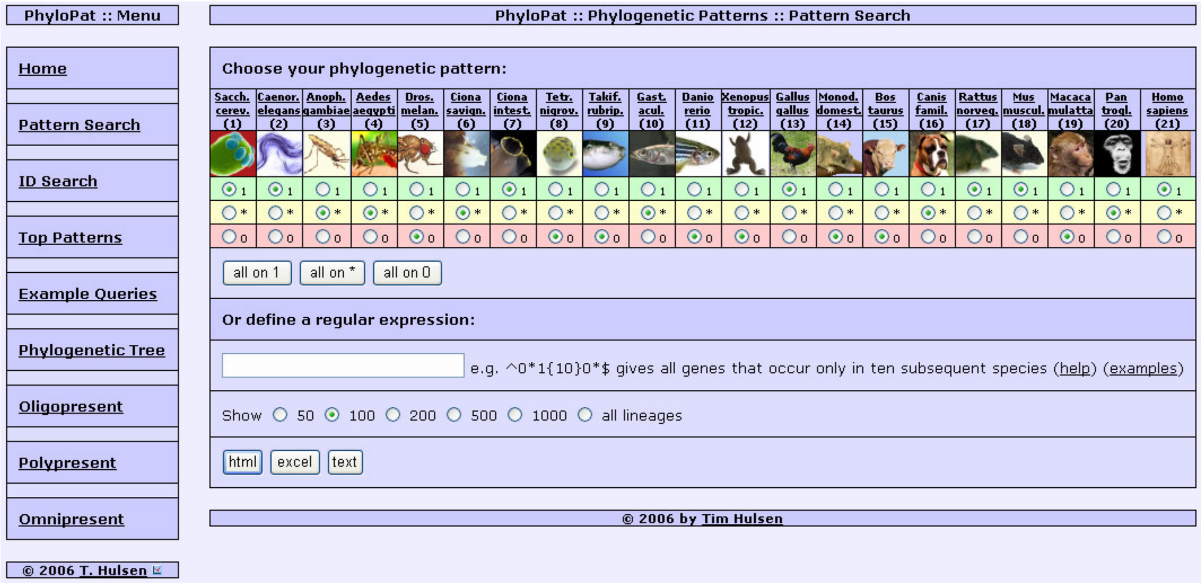
Chapter 3, figure 6. ROC_{50} and mean AP values for proteins larger than 500 aa

The ROC_{50} scores are shown at the left half, the mean AP values on the right half. Red bars: calculated using the Paracel Smith-Waterman algorithm. Blue bars: calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Green bars: calculated using the BLAST algorithm. Yellow bars: calculated using the FASTA algorithm. Purple bars: calculated using the SSEARCH algorithm. Orange bars: calculated using the ParAlign Smith-Waterman algorithm.



Chapter 3, figure 7. ROC_{50} and mean AP values for the SW scores of three different SW algorithms.

The ROC_{50} scores are shown at the left half, the mean AP values on the right half. Blue bars: calculated using the Biofacet Smith-Waterman algorithm with Z-score statistics. Purple bars: calculated using the SSEARCH algorithm. Orange bars: calculated using the ParAlign Smith-Waterman algorithm.



PhyloPat :: Menu

- Home
- Pattern Search
- ID Search
- Top Patterns
- Example Queries
- Phylogenetic Tree
- Oligopresent
- Polypresent
- Omnipresent

© 2006 T. Hulsen

PhyloPat :: Phylogenetic Patterns :: Pattern Search

Choose your phylogenetic pattern:

Sacch. cerev. (1)	Caenor. elegans (2)	Anoph. gambiae (3)	Aedes aegypti (4)	Dros. melan. (5)	Ciona savign. (6)	Ciona intest. (7)	Tetr. nigrov. (8)	Takif. rubrip. (9)	Gast. acul. (10)	Danio rerio (11)	Xenopus tropic. (12)	Gallus gallus (13)	Monod. domest. (14)	Bos taurus (15)	Canis famil. (16)	Rattus norveg. (17)	Mus muscul. (18)	Macaca mulatta (19)	Pan trogl. (20)	Homo sapiens (21)
● 1	● 1	○ 1	○ 1	○ 1	○ 1	● 1	○ 1	○ 1	○ 1	○ 1	○ 1	● 1	○ 1	○ 1	○ 1	● 1	● 1	○ 1	○ 1	● 1
○ *	○ *	● *	● *	○ *	● *	○ *	○ *	○ *	● *	○ *	○ *	○ *	○ *	○ *	● *	○ *	○ *	○ *	● *	○ *
○ 0	○ 0	○ 0	○ 0	● 0	○ 0	○ 0	● 0	● 0	○ 0	● 0	● 0	○ 0	● 0	● 0	○ 0	○ 0	○ 0	● 0	○ 0	○ 0

Or define a regular expression:

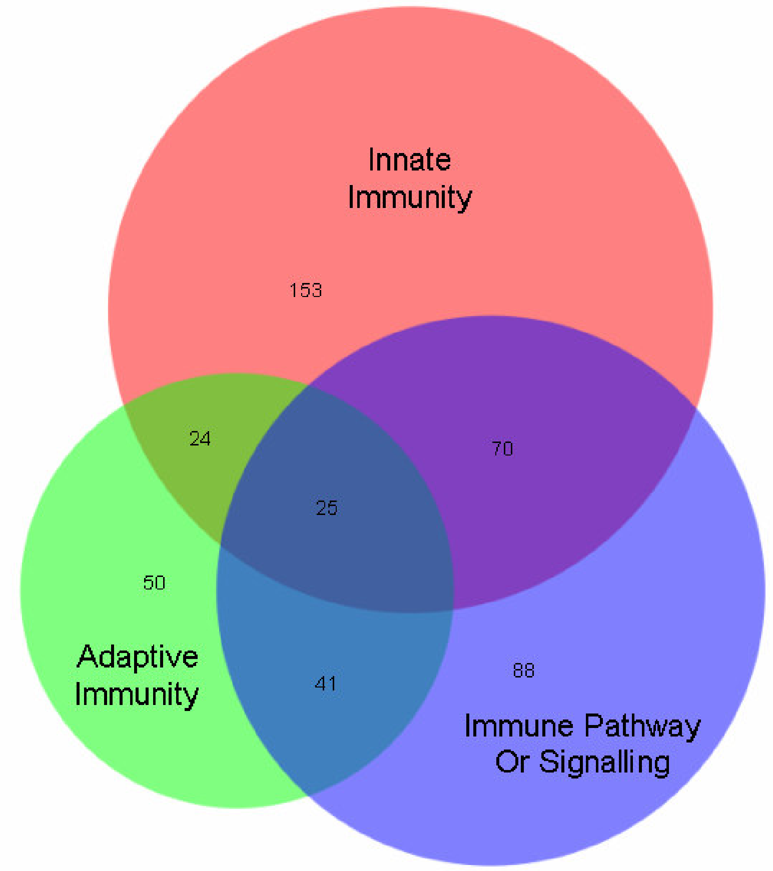
e.g. ^0*1{10}0*$ gives all genes that occur only in ten subsequent species (help) (examples)

Show ○ 50 ● 100 ○ 200 ○ 500 ○ 1000 ○ all lineages

© 2006 by Tim Hulsen

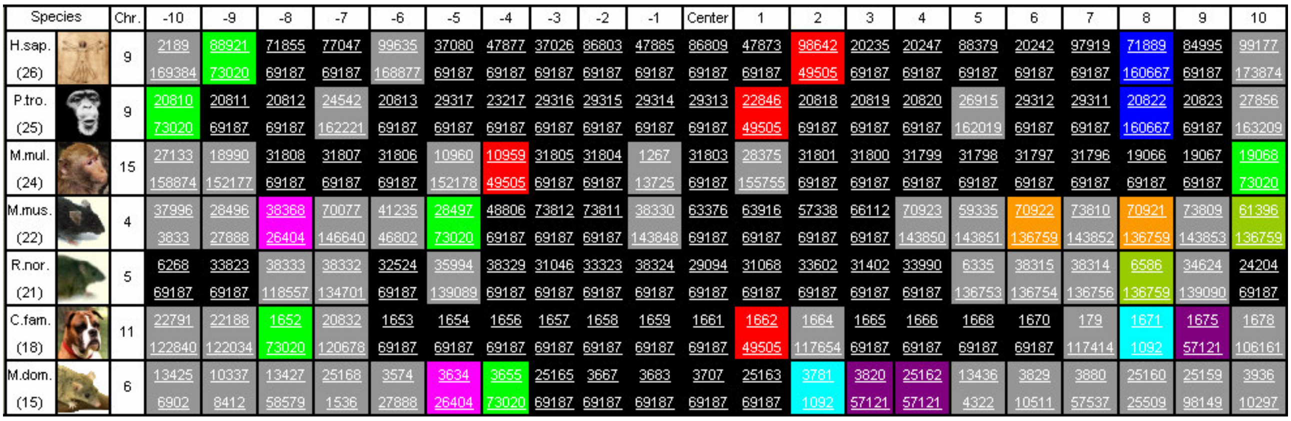
Chapter 4, figure 3. The PhyloPat web interface (Pattern Search tab)

The web interface has the menu on the left and the input/results page on the right. On the pattern search page, the user can generate a phylogenetic pattern by clicking a radio button for each species. 1 = present, * = present/absent, 0 = absent. The buttons directly below put all 21 species on the corresponding mode. MySQL regular expressions offer the possibility of advanced querying. The user can choose to show any number of lineages and choose the output format: HTML, Excel or plain text.



Chapter 5, figure 3. Venn diagram of the numbers of phylogenetic lineages linked to specific immunologic categories

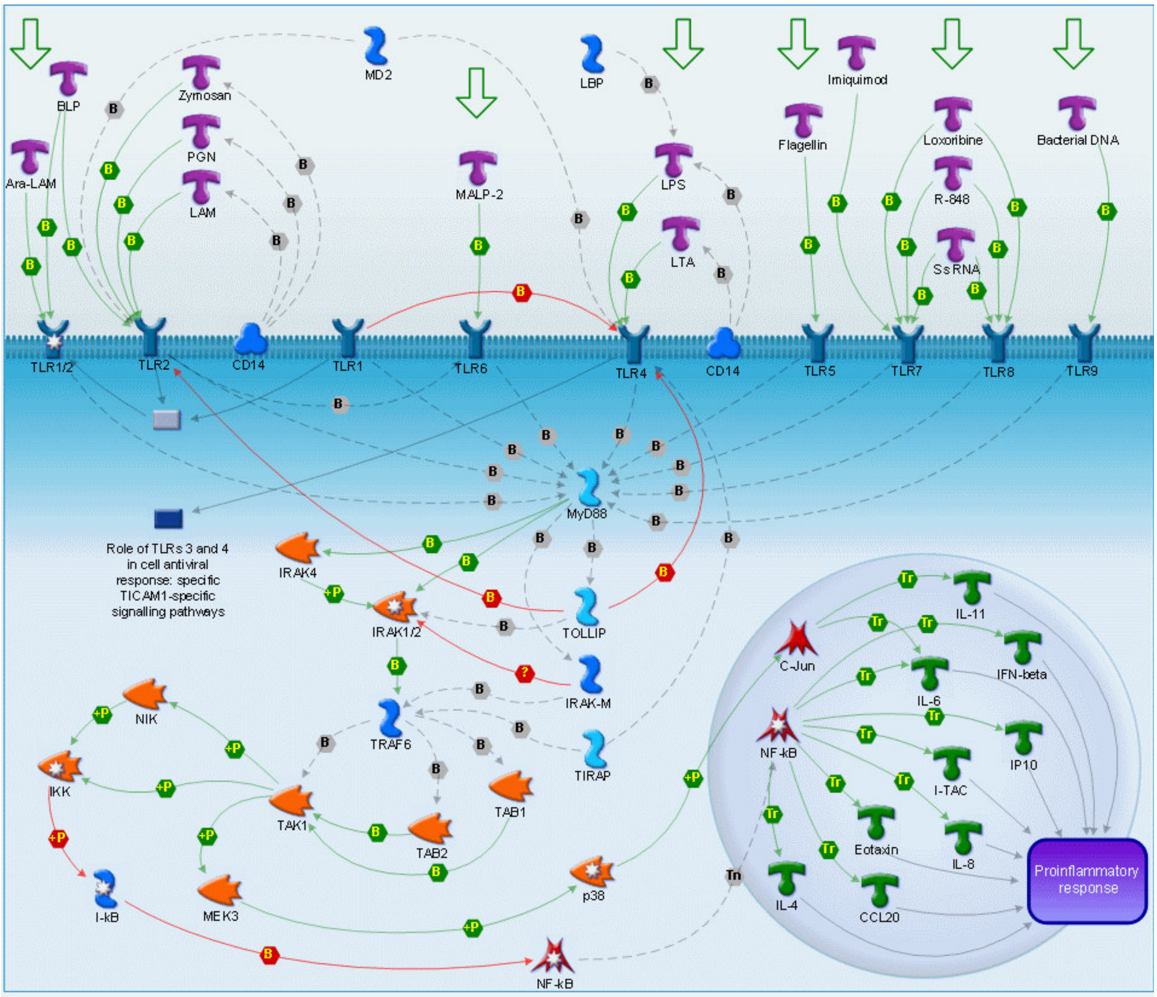
Venn diagram of the numbers of phylogenetic lineages linked to ‘Innate Immunity’ (red), ‘Adaptive Immunity’ (green) and ‘Immune Pathway or Signalling’ (blue) and combinations of these three categories. Each surface is proportional to the number it represents, except for the overlap between all three categories.



Species	Chr.	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	Center	1	2	3	4	5	6	7	8	9	10
H.sap. (26)	9	2189 169384	88921 73020	71855 69187	77047 69187	99635 168877	37080 69187	47877 69187	37026 69187	86803 69187	47885 69187	86809 69187	47873 69187	98642 49505	20235 69187	20247 69187	88379 69187	20242 69187	97919 69187	71889 160667	84995 69187	99177 173874
P.tro. (25)	9	20810 73020	20811 69187	20812 69187	24542 162221	20813 69187	29317 69187	23217 69187	29316 69187	29315 69187	29314 69187	29313 69187	22846 49505	20818 69187	20819 69187	20820 69187	26915 162019	29312 69187	29311 69187	20822 160667	20823 69187	27856 163209
M.mul. (24)	15	27133 158874	18990 152177	31808 69187	31807 69187	31806 69187	10960 152178	10959 49505	31805 69187	31804 69187	1267 13725	31803 69187	28375 155755	31801 69187	31800 69187	31799 69187	31798 69187	31797 69187	31796 69187	19066 69187	19067 69187	19068 73020
M.mus. (22)	4	37996 3833	28496 27888	38368 26404	70077 146640	41235 46802	28497 73020	48806 69187	73812 69187	73811 69187	38330 143848	63376 69187	63916 69187	57338 69187	66112 69187	70923 143850	59335 143851	70922 136759	73810 143852	70921 136759	73809 143853	61396 136759
R.nor. (21)	5	6268 69187	33823 69187	38333 118557	38332 134701	32524 69187	35994 139089	38329 69187	31046 69187	33323 69187	38324 69187	29094 69187	31068 69187	33602 69187	31402 69187	33990 69187	6335 136753	38315 136754	38314 136756	6586 136759	34624 139090	24204 69187
C.fam. (18)	11	22791 122840	22188 122034	1652 73020	20832 120678	1653 69187	1654 69187	1656 69187	1657 69187	1658 69187	1659 69187	1661 69187	1662 49505	1664 117654	1665 69187	1666 69187	1668 69187	1670 69187	179 117414	1671 1092	1675 57121	1678 106161
M.dom. (15)	6	13425 6902	10337 8412	13427 58579	25168 1536	3574 27888	3634 26404	3655 73020	25165 69187	3667 69187	3683 69187	3707 69187	25163 69187	3781 1092	3820 57121	25162 57121	13436 4322	3829 10511	3880 57537	25160 25509	25159 98149	3936 10297

Chapter 5, figure 4. Conservation of gene order for phylogenetic lineage IP377 (IFNA)

Conservation of gene order for phylogenetic lineage IP377 or PP069187, which consists of several members of the IFNA (interferon alpha) family, for seven species: *H. sapiens*, *P. troglodytes*, *M. mulatta*, *M. musculus*, *R. norvegicus*, *C. familiaris* and *M. domestica*. For each species, the most central IFNA gene is shown next to its twenty surrounding genes on the chromosome. Black: gene belonging to the IFNA phylogenetic lineage. Color: gene belonging to phylogenetic lineage with two or more members in this figure. Grey: belonging to phylogenetic lineages with only one member in this figure ('singleton'). Only the final five/six characters of each Ensembl ID or PPID are shown.



Chapter 5, figure 5. The toll-like receptor pathway

The pathway 'Toll-like receptor (TLR) ligands and common TLR signalling pathway leading to cell proinflammatory response' from the GeneGo MetaCore™ [28] application.

List of publications

Heavier-than-air flying machines are impossible
L. Oliveira, **T. Hulsen**, D. Lutje Hulsik, A.C.M. Paiva and G. Vriend
FEBS Letters 2004, 564 (3): 269-273
PubMed ID 15111108, doi:10.1016/S0014-5793(04)00320-5
http://www.febsletters.org/article/PIIS0014579304003205

Benchmarking ortholog identification methods using functional genomics data
T. Hulsen, M.A. Huynen, J. de Vlieg and P.M.A. Groenen
Genome Biology 2006, 7 (4): R31
PubMed ID 16613613, doi:10.1186/gb-2006-7-4-r31
http://genomebiology.com/2006/7/4/R31

PhyloPat: phylogenetic pattern analysis of eukaryotic genes
T. Hulsen, J. de Vlieg and P.M.A. Groenen
BMC Bioinformatics 2006, 7 (1): 398
PubMed ID 16948844, doi:10.1186/1471-2105-7-398
http://www.biomedcentral.com/1471-2105/7/398

Testing statistical significance scores of sequence comparison methods using structure similarity
T. Hulsen, J. de Vlieg, J.A.M. Leunissen and P.M.A. Groenen
BMC Bioinformatics 2006, 7 (1); 444
PubMed ID 17038163, doi: 10.1186/1471-2105-7-444
http://www.biomedcentral.com/1471-2105/7/444

Identification of novel functional TBP-binding sites and general factor repertoires
S. Denissov, M.A. van Driel, R. Voit, M.L. Hekkelman, **T. Hulsen**, N. Hernandez, I. Grummt, R. Wehrens and H.G. Stunnenberg
EMBO J. 2007
PubMed ID 17268553, doi: 10.1038/sj.emboj.7601550
http://www.nature.com/emboj/journal/vaop/ncurrent/full/7601550a.html

The construction of genome-based transcriptional units
S. van Hooff, R. Monajemi, J. Koster, **T. Hulsen**, M. Roos, B.D.C. van Schaik, M.F. van Batenburg, R. Versteeg and A.H.C. van Kampen
Submitted

Dynamics of head-to-head genes in vertebrates
E. Franck, **T. Hulsen**, M.A. Huynen, N.H. Lubsen, W.W. de Jong and O. Madsen

Manuscript in preparation

Evolution of the immune system from model organism to man
T. Hulsen, W.W.M. Fleuren, H.H.D. Kerstens, M.A.M. Groenen and P.M.A. Groenen
Manuscript in preparation

List of conferences

2002-06-14	Wageningen (NL)	Bioinformatics 2002: 'The Best Of Both Worlds'
2002-09-17	Utrecht (NL)	Symposium Comparative Genomics
2002-12-04	The Hague (NL)	Genomics Momentum 2002
2002-12-12	Amsterdam (NL)	SARA Superdag (Oral presentation)
2003-03-06 - 2003-03-07	Schoorl (NL)	NVTB Meeting
2003-04-28 - 2003-04-29	Nijmegen (NL)	NSRIM Symposium
2003-05-08 - 2003-05-09	Arnhem (NL)	ICS PhD Two-Day Conference 2003 (Poster presentation)
2003-06-20	Utrecht (NL)	Bioinformatics 2003: 'Bioinformatics at the Interface' (Poster presentation)
2003-09-04	Amsterdam (NL)	BioASP Users Forum: ‘Protein World/Biofacet’
2003-09-27 - 2003-09-30	Paris (F)	ECCB 2003 (Poster presentation)
2003-10-29	Ravenstein (NL)	CMBI Conference (Poster presentation)
2003-11-05	Oss (NL)	NVBMB Najaarssymposium 2003: 'Protein-protein interactions in cells: the biochemistry of signalling' (Poster presentation)
2003-11-15 - 2003-11-21	Phoenix (USA)	SC2003: 'Igniting Innovation' (Poster presentation)
2003-11-27	Amsterdam (NL)	Genomics Momentum 2003 (Poster presentation)
2004-01-20	Utrecht (NL)	GeNeYouS Symposium 2004: 'Decisions in Genomics'
2004-03-22 - 2004-03-23	Amsterdam (NL)	First International Symposium on networks in Bioinformatics
2004-04-19 - 2004-04-20	Arnhem (NL)	ICS PhD Two-Day Conference 2004

		(Poster presentation)
2004-07-31 - 2004-08-04	Glasgow (UK)	ISMB/ECCB 2004 (Poster presentation)
2004-08-30 - 2004-09-01	Rotterdam (NL)	Genomics Momentum 2004: ‘Genomics for Our World’
2004-11-30	Utrecht (NL)	Koppelevenement Bioinformatica
2004-12-09	Amsterdam (NL)	BioASP Users Forum: 'Pathways'
2005-04-11 - 2005-04-15	Lyon (F)	BioVision 2005
2005-04-27	Oss (NL)	GeNeYouS Symposium 2005 (Organization + Poster presentation)
2005-04-28 - 2005-04-29	Arnhem (NL)	ICS PhD Two-Day Conference 2005 (Organization + Poster presentation)
2005-05-12	Utrecht (NL)	Bio Career Event 2005 (Organization)
2005-09-28 - 2005-10-01	Madrid (E)	ECCB 2005
2005-10-05	The Hague (NL)	Genomics Momentum 2005: ‘Genomics in the here and now: Emerging benefits for health’
2006-04-24	Ede (NL)	Netherlands Bioinformatics Conference (Poster presentation)
2006-04-27 - 2006-04-28	Arnhem (NL)	ICS PhD Two-Day Conference 2006 (Organization + Oral presentation)
2006-10-17 - 2006-10-18	Wageningen (NL)	BBC 2006: ‘Bioinformatics for Food and Health’ (Oral presentation)
2006-11-09	Rotterdam (NL)	Genomics Momentum 2006: ‘Genomics: Ready for the next step’
2007-01-09	Nijmegen (NL)	Bioinformatics Symposium
2007-02-25 – 2007-02-27	Enschede (NL)	ELSYS 2007 (Poster presentation)

Glossary

Adverse effect

An abnormal, harmful, undesired and/or unintended side-effect, although not necessarily unexpected, which is obtained as a result of a therapy or other medical intervention, such as drug/chemotherapy, physical therapy, surgery, medical procedure, use of a medical device, etc.

Amino acid

A molecule that contains both amine and carboxylic acid functional groups. They are the basic structural building units of proteins. They form short polymer chains called peptides or polypeptides which in turn form structures called proteins.

Analogy

Two proteins or genes are said to be analogous if they perform the same or similar function by a similar mechanism. These similar mechanisms may have evolved through different pathways, a process known as convergent evolution. The concept of analogy is contrasted with that of homology.

Annotation

The connection of a previously unknown sequence representation of genetic material with information relating position to intron-exon-boundaries, regulatory sequences, repeats, gene names and protein products, etc.. This annotation is usually stored in predefined fields in biological databases, especially sequence databases.

Bacteriophage

A virus that infects bacteria. The term is commonly used in its shortened form, phage.

BLAST

Basic Local Alignment Search Tool. An algorithm for comparing biological sequences, such as the amino-acid sequences of different proteins or the DNA sequences.

Comparative genomics

The study of relationships between the genomes of different species or strains. Comparative genomics is an attempt to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes.

Drug discovery

In medicine, biotechnology and pharmacology, drug discovery is the process by which drugs are discovered and/or designed. The process of drug discovery involves the identification of candidates, synthesis, characterization, screening, and assays for therapeutic efficacy. Once a compound has shown its value in these tests, it will begin the process of drug development prior to clinical trials.

Dynamic programming

A method for reducing the runtime of algorithms exhibiting the properties of overlapping subproblems and optimal substructure.

Eukaryote

An organism with a complex cell or cells, in which the genetic material is organized into a membrane-bound nucleus or nuclei. Eukaryotes comprise animals, plants, and fungi—which are mostly multicellular—as well as various other groups that are collectively classified as protists (many of which are unicellular).

Expression

The process by which a gene's DNA sequence is converted into the structures and functions of a cell. Gene expression is a multi-step process that begins with transcription of DNA, which genes are made of, into messenger RNA. It is then followed by post transcriptional modification and translation into a gene product, followed by folding, post-translational modification and targeting.

FASTA A DNA and Protein sequence alignment software package first described (as FASTP) by David J. Lipman and William R. Pearson in 1985. The original FASTP program was designed for protein sequence similarity searching. FASTA, described in 1988, added the ability to do DNA:DNA searches, translated protein:DNA searches and provided a more sophisticated shuffling program for evaluating statistical significance.

Genome The whole hereditary information of an organism that is encoded in the DNA (or, for some viruses, RNA). This includes both the genes and the non-coding sequences.

Genomics The study of an organism's genome and the use of the genes. It deals with the systematic use of genome information, associated with other data, to provide answers in biology, medicine, and industry.

G protein-coupled rec. Also known as seven transmembrane receptors, 7TM receptors, and heptahelical receptors. A protein family of transmembrane receptors that transduce an extracellular signal (ligand binding) into an intracellular signal (G protein activation). The GPCRs are the largest protein family known, members of which are involved in all types of stimulus-response pathways, from intercellular communication to physiological senses.

Hemoglobin The iron-containing oxygen-transport metalloprotein in the red cells of the blood in mammals and other animals. Hemoglobin in vertebrates transports oxygen from the lungs to the rest of the body, such as to the muscles, where it releases the oxygen load. Hemoglobin also has a variety of other gas-transport and effect-modulation duties, which vary from species to species, and which in invertebrates may be quite diverse.

Homeobox A DNA sequence found within genes that are involved in the regulation of development (morphogenesis) of animals, fungi and plants. Genes that have a homeobox are called homeobox genes and form the homeobox gene family.

Homology Is used in reference to protein or DNA sequences, meaning that the given sequences share a common ancestor. Sequence homology may also indicate common function.

Hox genes A particular subgroup of homeobox genes, that are found in a special gene cluster, the Hox cluster (also called Hox complex). Hox genes function in patterning the body axis. Thus, by providing the identity of particular body regions, Hox genes determine where limbs and other body segments will grow in a developing fetus or larva.

Interleukins A group of cytokines that were first seen to be expressed by white blood cells (leukocytes, hence the -leukin) as a means of communication (inter-). The name is sort of a relic though; it has since been found that interleukins are produced by a wide variety of bodily cells. The function of the immune system depends in a large part on interleukins, and rare deficiencies of a number of them have been described, all featuring autoimmune diseases or immune deficiency.

Ligand An atom, ion, or molecule that generally donates one or more of its electrons through a coordinate covalent bond to, or shares its electrons through a covalent bond with one or more central atoms or ions.

Model organism A species that is extensively studied to understand particular biological phenomena, with the expectation that discoveries made in the organism model will provide insight into the workings of other organisms (e.g. humans). This is possible because fundamental biological principles such as metabolic, regulatory, and developmental pathways, and the genes that code for them, are conserved through evolution.

Multiple seq. alignment A sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In general, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins.

Nucleic acid A complex, high-molecular-weight biochemical macromolecule composed of nucleotide chains that convey genetic information. The most common nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Nucleic acids are found in all living cells and viruses.

Nucleotide A chemical compound that consists of a heterocyclic base, a sugar, and one or more phosphate groups. In the most common nucleotides the base is a derivative of purine or pyrimidine, and the sugar is the pentose (five-carbon sugar) deoxyribose or ribose.

Oligopresent genes Genes that are present in only one or two species within a selected species set.

Omnipresent genes Genes that are present in all species within a selected species set.

Orthology Homologous sequences are orthologous if they were separated by a speciation event: if a gene exists in a species, and that species diverges into two species, then the copies of this gene in the resulting species are orthologous.

Paralogy Homologous sequences are paralogous if they were separated by a gene duplication event: if a gene in an organism is duplicated, then the two copies are paralogous.

Pharmacogenomics The branch of pharmaceutics which deals with the influence of genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with a drug's efficacy or toxicity.

Phylogenetic tree A phylogenetic tree is a tree showing the evolutionary interrelationships among various species or other entities, such as genes or proteins, that are believed to have a common ancestor. A hylogenetic tree is a form of a cladogram. In a phylogenetic tree, each node with descendants represents the most recent common ancestor of the descendants, and edge lengths correspond to time estimates.

Phylogenomics A method of assigning a function to a gene based on its evolutionary history in a Phylogenetic tree. Phylogenomics uses knowledge on the evolution of a gene to improve function prediction.

Polypresent genes Genes that are present in almost all species within a selected species set.

Prokaryote
An organism without a cell nucleus (= karyon), or indeed any other membrane-bound organelles, in most cases unicellular (in rare cases, multicellular).

Proteome
The entire complement of proteins in a given biological organism or system at a given time, i.e. the protein products of the genome.

Sequence alignment
A way of arranging DNA, RNA, or protein primary sequences to emphasize their regions of similarity, which may indicate functional or evolutionary relationships between the genes or proteins in the query. Sequences are typically written with their characters (generally amino acids or nucleotides) in aligned columns that into which gaps are inserted so that successive columns contain identical or similar characters.

Sequencing
The process of determining the nucleotide order of a given DNA fragment. Currently, almost all DNA sequencing is performed using the chain termination method developed by Frederick Sanger. This technique uses sequence-specific termination of an DNA synthesis reaction using modified nucleotide substrates.

Smith-Waterman
A well-known algorithm for performing local sequence alignment; that is, for determining similar regions between two nucleotide or protein sequences.

Substrate
A molecule upon which an enzyme acts. Enzymes catalyze chemical reactions involving the substrate(s). The substrate binds with the enzyme's active site, and an enzyme-substrate-complex is formed. The substrate is broken down into a product and is released from the active site. The active site is now free to accept another substrate molecule.

Toll-like receptors
Type I transmembrane proteins that recognize pathogens and activate immune cell responses as a key part of the innate immune system. In vertebrates, they can help activate the adaptive immune system, linking innate and acquired immune responses.

Transcription
The process through which a DNA sequence is enzymatically copied by an RNA polymerase to produce a complementary RNA. Or, in other words, the transfer of genetic information from DNA into RNA.

Translation
The second process of protein biosynthesis (part of the overall process of gene expression). In translation, Messenger RNA (mRNA) is decoded to produce a specific polypeptide according to the rules specified by the genetic code. Translation is necessarily preceded by transcription.

Vertebrate
A subphylum of chordates, specifically, those with backbones or spinal columns. About 58,000 species of vertebrates have been described.

Xenology
Homology that arises via horizontal gene transfer between unrelated species

Summary

The current drug discovery pipeline can be regarded as slow and inefficient. The average time spent for the complete process is around fourteen years, and the 'attrition rate', i.e. the ratio tested compounds / approved drugs, is currently almost 99%. In this thesis, we discuss a pharmacophylogenomics approach to shorten and improve this pipeline.

In **chapter 1**, we give an introduction to comparative genomics, drug discovery, and their intersection of pharmacophylogenomics. It introduces chapters 2-4 as general orthology/genomics methodologies, and chapter 5-7 as applications of orthology in, respectively, immunology, evolutionary biology and transcriptomics.

The orthology benchmark described in **chapter 2** uses the basic assumption that orthologs have a highly similar function. This property is important because it forms the basis for the main goal of ortholog identification: the transfer of functional annotation of proteins from one species to the other. We use functional genomics data as a benchmark for a number of ortholog identification methods and we show that there is a trade-off between sensitivity and selectivity of the several methods. This means that methods such as the best bidirectional hit method, which only give a small number of orthologous pairs, are best in predicting functional similarities. Methods such as COG (Clusters of Orthologous Groups) give a large number of orthologous pairs, but are not as good as other methods in predicting functional similarities. We tried to combine the two factors of sensitivity and selectivity into one score and concluded that the InParanoid method is the best ortholog identification method.

The quality of an ortholog identification method depends not only on the method itself, but also on its settings and other programs used in the process. The InParanoid program for example, needs as input a list of sequence pairs together with their similarity scores, provided by a sequence comparison algorithm such as BLAST. **Chapter 3** shows us that BLAST could better be replaced by other sequence comparison algorithms: SSEARCH or ParAlign. These methods are usually slower than BLAST, but for most large sequence comparison projects time is not the limiting factor. Improvements could also be made by using different settings than the defaults, or different statistical significance values. We recommend e-value and not z-score, because of its high accuracy and calculation speed.

Chapter 4 shows how orthology information can be used to create phylogenetic patterns, which display the presence or absence of certain genes over a set of species. They can be used, for example, to cluster genes that occur in the same species or taxons. These genes are likely to have a similar function or to be involved in the same biological process. They can also be used to study gene families and their expansions or deletions over time. Another use lies in the study of anti-correlating patterns: genes that have exactly inversed patterns. These genes might be completely different in function, but could also be analogous: performing the same function in different species or taxons, without having a common ancestor. The PhyloPat tool is not only useful for phylogenetic pattern querying; it offers the functionality for all kinds of evolutionary studies, and can be used

for the annotation of proteins with unknown function. This functional annotation is one of the most important direct applications of functional genomics.

The main application of orthology in drug discovery lies in the assumption that differences in drug response between human and model organisms can be explained by looking at the orthologous proteins between these species. In this thesis we tried to show that single proteins, and even protein families or complete protein pathways, can be linked cross-species by identifying orthologous relationships. The study of the evolution of the immune system from model organisms (chicken, rat, mouse, etc.) to man, described in **chapter 5**, is a good example of this. For drug discovery the interspecies mapping of protein pathways is of specific interest. A different response to a certain drug in man and in a model organism can be elucidated by mapping the organisms' pathways onto each other. This could increase the predictive value of studies in animal models drastically. However, studies like this need highly accurate and reliable orthology information. Moreover, other factors like alternative transcripts, expression levels and three-dimensional structure could be part of the solution. All in all, in order to provide an answer to pharmacogenomics questions, a whole range of genomics data might be needed, instead of just orthology data.

The concept of orthology has many applications, and not only in drug discovery. In **chapter 6** we have shown that it can be used for evolutionary studies, in this case the evolution of bidirectional gene pairs. Using Ensembl orthologies we were able to map bidirectional (head-to-head) gene pairs from species A to gene pairs from species B. In this way we connected gene pairs from eleven vertebrate species to each other. We found an enrichment of head-to-head gene pairs with distance less than 600 bp in the human, chimpanzee, mouse, rat and chicken genomes, and an enrichment of head-to-tail gene pairs in fish and *Ciona*. We concluded that this indicates a transition from head-to-tail gene orientation in lower vertebrates to head-to-head orientation in higher vertebrates. A study like this would not have been possible without complete and accurate orthology information.

Another application of orthology is shown in **chapter 7**. Here we discussed the construction of transcriptional units, i.e. groups of EST and mRNA sequences that actually belong to one single gene, for every organism in the UCSC genome database. The human and mouse TU sets were mapped onto each other using ortholog identification methods, enabling us to compare cross-species data. In the future, this methodology will be used to map TU sets from more species. The quality of this mapping, and the conclusions drawn from it, are largely dependent on the accuracy of the ortholog prediction.

Chapter 8 concludes this thesis with the statement that application of orthology alone is not likely to answer many research questions: a wide range of functional genomics data needs to be gathered to shed light on complex systems such as pathways in the field of drug discovery. However, the completeness and higher reliability of future genomics data will enable researchers to perform studies like in this thesis in more detail and with more accuracy. This thesis offers a good foundation for this future pharmacophylogenomics research.

Samenvatting

De huidige drug discovery pipeline kan worden beschouwd als langzaam en inefficiënt. De gemiddelde tijdsduur voor het complete proces ligt rond de veertien jaar, en de 'attrition rate', oftewel de ratio geteste compounds / goedgekeurde medicijnen, is momenteel bijna 99%. In dit proefschrift bediscussiëren we een pharmacophylogenomics aanpak om deze pipeline te verkorten en te verbeteren.

In **hoofdstuk 1** geven we een inleiding op comparative genomics, drug discovery, en hun intersectie genaamd pharmacophylogenomics. Het introduceert de hoofdstukken 2-4 als algemene orthologie/genomics methodologieën, en hoofdstuk 5-7 als toepassingen van orthologie in respectievelijk immunologie, evolutionaire biologie en transcriptomics.

De orthologie benchmark beschreven in **hoofdstuk 2** gebruikt de basisveronderstelling dat orthologen een zeer similaire functie hebben. Deze eigenschap is belangrijk omdat het de basis vormt voor het hoofddoel van ortholoog-identificatie: het overbrengen van de functionele annotatie van eiwitten van de ene soort naar de andere. We gebruiken functional genomics gegevens als een benchmark voor een aantal ortholoog-identificatie methodes en we laten zien dat er een wisselwerking is tussen de sensitiviteit en selectiviteit van de verscheidene methodes. Dit betekent dat methodes als de 'best bidirectional hit' methode, die slechts een klein aantal orthologe paren geeft, het best zijn in het voorspellen van functionele similariteiten. Methodes als COG (Clusters of Orthologous Groups) geven een groot aantal orthologe paren, maar zijn niet zo goed als andere methodes in het voorspellen van functionele similariteiten. Wij hebben geprobeerd de twee factoren van sensitiviteit en selectiviteit te combineren in één enkele score en hebben geconcludeerd dat de InParanoid methode de beste ortholoog-identificatie methode is.

De kwaliteit van een ortholoog-identificatie methode hangt niet alleen af van de methode zelf, maar ook van zijn instellingen en van andere programma's die zijn gebruikt in het proces. Het InParanoid programma bijvoorbeeld, gebruikt als input een lijst van sequentie-paren plus hun similariteitsscores, verschaft door een sequentievergelijkings-algoritme zoals BLAST. **Hoofdstuk 3** laat ons zien dat BLAST beter vervangen kan worden door andere sequentievergelijkings-algoritmes: SSEARCH of ParAlign. Deze methodes zijn gebruikelijk langzamer dan BLAST, maar voor de meeste grootschalige sequentievergelijkings-projecten is tijd niet de limiterende factor. Verbeteringen kunnen ook worden verkregen door andere instellingen te gebruiken dan de standaard instellingen, of andere statistische significantie waarden. Wij bevelen de e-value en niet de z-score aan, vanwege zijn hogere nauwkeurigheid en berekeningssnelheid.

Hoofdstuk 4 toont aan hoe orthologie informatie kan worden gebruikt om phylogenetische patronen te creëren, die de aanwezigheid of afwezigheid van bepaalde genen over een aantal soorten weergeven. Ze kunnen bijvoorbeeld worden gebruikt om genen te clusteren die in dezelfde soorten of taxa voorkomen. Deze genen hebben waarschijnlijk een similaire functie of zijn betrokken bij hetzelfde biologische proces. Ze kunnen ook worden gebruikt om genfamilies en hun expansies of deleties te bestuderen. Een andere toepassing ligt in de studie van anti-correlerende patronen: genen die exact omgekeerde patronen hebben. Deze genen kunnen een compleet verschillende functie hebben, maar ze kunnen ook analoog zijn, oftewel ze kunnen dezelfde functie vervullen in verschillende soorten of taxa, zonder dat ze een gemeenschappelijk voorouder-gen hebben. De PhyloPat applicatie is niet alleen bruikbaar voor het zoeken met behulp van phylogenetische patronen; het heeft functionaliteit voor allerlei soorten evolutionaire studies, en kan worden gebruikt voor de annotatie van eiwitten met onbekende functies. Deze functionele annotatie is één van de meest belangrijke directe toepassingen van functional genomics.

De hoofdtoepassing van orthologie in drug discovery ligt in de veronderstelling dat verschillen in drug respons tussen mens en model-organismen verklaard kunnen worden door te kijken naar de orthologe eiwitten tussen deze soorten. In dit proefschrift hebben we geprobeerd aan te tonen dat eiwitten, en zelfs eiwit-families of complete eiwit-pathways, cross-species gelinkt kunnen worden door orthologe relaties te identificeren. De studie van de evolutie van het immuunsysteem van model-organismen (kip, rat, muis, etc.) naar mens, beschreven in **hoofdstuk 5**, is een goed voorbeeld hiervan. Voor drug discovery is de interspeciële mapping van eiwit-pathways van specifiek belang. Een verschillende respons op een bepaalde drug in mens en in een model-organisme kan worden verduidelijkt door de pathways van beide organismen over elkaar te leggen. Dit zou de voorspellende waarde van studies in diermodellen drastisch kunnen verhogen. Echter, studies als deze hebben zeer nauwkeurige en betrouwbare orthologie informatie nodig. Bovendien zouden andere factoren zoals alternatieve transcripten, expressie-niveaus en driedimensionale structuur een deel van de oplossing kunnen vormen. Alles bij elkaar, om een oplossing te vinden voor pharmacogenomics kwesties is mogelijk een hele verzameling genomics data nodig, in plaats van alleen orthologie data.

Het concept van orthologie heeft vele toepassingen, en niet alleen in drug discovery. In **hoofdstuk 6** hebben we laten zien dat het gebruikt kan worden voor evolutionaire studies, in dit geval de evolutie van bidirectionele genparen. Met behulp van Ensembl orthologiëen waren we in staat om bidirectionele (head-to-head) genparen uit soort A op genparen uit species B te mappen. Op deze manier verbonden we genparen uit elf vertebrate soorten met elkaar. We vonden een verrijking van head-to-head genparen met een afstand van minder dan 600 bp in het mensen-, chimpanseëen, muizen-, ratten- en kippen-genoom, en een verrijking van head-to-tail genparen in de vis en *Ciona*. We concluderen dat dat duidt op een overgang van head-to-tail gen-orientatie in lagere vertebraten naar een head-to-head gen-orientatie in hogere vertebraten. Een studie als deze zou niet mogelijk zijn geweest zonder complete en nauwkeurige orthologie informatie.

Nog een toepassing van orthologie is te zien in **hoofdstuk 7**. Hier behandelden we de constructie van ‘transcriptional units’ (TU), oftewel groepen van EST en mRNA sequenties die feitelijk behoren tot één enkel gen, voor elk organisme in de UCSC genoom database. De TU sets uit mens en muis werden over elkaar gelegd met behulp van ortholoog identificatie methodes, wat ons in staat stelde data uit verschillende soorten te vergelijken. In de toekomst zal deze methodologie gebruikt worden om TU sets uit meerdere soorten over elkaar te leggen. De kwaliteit van deze mapping, en de conclusies die eruit worden getrokken, zijn grotendeels afhankelijk van de nauwkeurigheid van de orthologie-voorspelling.

Hoofdstuk 8 besluit dit proefschrift met de bewering dat het toepassen van orthologie alleen waarschijnlijk weinig onderzoeksvragen zal beantwoorden: een grote verscheidenheid aan functional genomics data moet worden verzameld om duidelijkheid te verschaffen over complexe systemen als pathways in het veld van de drug discovery. Echter, de volledigheid en grotere betrouwbaarheid van toekomstige genomics gegevens zal onderzoekers in staat stellen om studies zoals beschreven in dit proefschrift met meer detail en meer nauwkeurigheid uit te voeren. Dit proefschrift biedt een goede basis voor dit toekomstig pharmacophylogenomics onderzoek.

Dankwoord / Acknowledgements

En dan nu … het meest gelezen onderdeel van het proefschrift: het dankwoord.

In de herfst van 2000 begon ik mijn stage bij het CMBI. In die tijd was deze afdeling een stuk kleiner dan nu. Gert, bedankt dat je David en mij wegwijs maakte in de GPCRs. David, het was altijd erg gezellig, bedankt daarvoor. We zijn nog steeds goede vrienden. Heel erg bedankt dat je mijn paranimf wil zijn. Ook mijn andere kamergenoot Jorn, inmiddels al vader, dankjewel voor de gezelligheid. Andere studenten in die periode waren Sander, Jos, Koen en (AiO) Simon.

Na mijn stage werkte ik ruim een jaar op uitzendbasis voor Organon. Uit die tijd wil ik ook een paar mensen met name bedanken. Jeroen, bedankt voor de eerste opzet van Protein World en succes met je eigen promotie. Blaise, ik hoop dat je wat aan mijn Python adviezen hebt gehad. Ik wil je vooral bedanken voor de prachtige vakantie in Kameroen, met name mogelijk gemaakt door je vrouw Liliane en de rest van de familie en vriendenkring. En succes met je eigen promotie, het zal niet lang meer duren.

Uit het eerste jaar promotie-onderzoek: Martijn, ondanks dat de samenwerking niet altijd even vlotjes verliep, toch bedankt voor de allereerste kennismaking met de wereld van comparative genomics, en je hulp bij mijn eerste artikel (hoofdstuk 2). Berend, bedankt voor je wetenschappelijke adviezen. Bas, bedankt voor de rood-groen-zwarte poef.

Twee personen uit de ‘Comics’ groep verdienen absoluut hun eigen alinea.
Guenola, omdat we ongeveer tegelijkertijd zijn begonnen, en met vergelijkbare problemen kampten, hebben we veel steun bij elkaar kunnen vinden. Bovendien was je net zo e-mail verslaafd als ik. Merci beaucoup et bonne chance, wat je ook gaat doen in de toekomst (wetenschap of schilderen).
Toni, ik herinner me vooral nog lange nachten in het uitgaanscentrum van Glasgow en Madrid, maar natuurlijk ook gewoon in Nijmegen. Het ga je goed amigo, maar dat zal wel lukken nu je, terug in hometown Valencia, vanaf je werkplek uitzicht hebt op de dolfijnenshow.

Ook buiten de Comics groep waren er een aantal mensen met wie ik veel contact had. Richard, leuk al die gesprekken over orthologie, voetbal en reizen (in willekeurige volgorde). Marc, bedankt voor je adviezen, zowel wetenschappelijk als anderszins. Er was altijd wat te zien in restaurant De Refter, of het nou op drie, zes of negen uur was.

Na ruim een jaar ging ik verder bij Organon. Uit deze tijd kwam het grootste gedeelte van het proefschrift voort. Peter, bedankt voor je begeleiding. Van alle personen die ik hier bedank is jouw bijdrage aan het proefschrift natuurlijk het grootst geweest. Jacob, uitstekend dat je in je drukke schema nog tijd hebt kunnen vinden om mijn promotor te kunnen zijn. René, ‘grote kleine baas’, dankjewel voor je carrière adviezen.

Wilco 'Hookipa', je was gedurende een erg lange tijd mijn kamergenoot, en mijn vaste 'Protein World' maatje. Bedankt voor je technische ondersteuning, bij met name het kippen-project (hoofdstuk 5). Heel fijn dat je zoveel van kippetjes afwist, het kwam goed van pas.

Ik wil ook graag iedereen uit andere onderzoeksgroepen bedanken met wie ik heb samengewerkt. Erik, Ole, ik heb erg prettig met jullie samengewerkt. De evolutie van head-to-head genes was een erg leuk project. Vooral bedankt dat jullie nog even extra de schouders eronder hebben gezet om hoofdstuk 6 op tijd af te krijgen. Jack, bedankt voor je tips bij het artikel over sequentievergelijking (hoofdstuk 3) en andere bioinformatica-gerelateerde zaken. Sander, Ramin, Antoine, ook jullie bedankt dat hoofdstuk 7 op tijd is afgerond. Allemaal veel succes bij het verdere onderzoek naar transcriptome maps. Sergei, Henk, onze samenwerking stamt nog uit de allereerste fase van mijn promotie-onderzoek. Fijn dat jullie mijn PromScan tool hebben kunnen gebruiken voor jullie ChIP-on-chip studie. Sergei, veel succes met je promotie. Hinri, Martien, ik hoop dat er alsnog wat moois voortkomt uit het kippenproject.

Ook was er ondersteuning op andere fronten. Bijvoorbeeld van het secretariaat, zowel op het CMBI als bij Organon. Barbara, Esther, Thea, Ria, jullie werkten altijd uitstekend mee als er wat geregeld moest worden. Maar ook de systeembeheerders en andere technische personen hebben me geweldig geholpen: Wim, Wilmar, Stefan, Tinka, Ruud. Plus alle andere collega's bij CMBI en Organon: iedereen bedankt!

Een voor mij heel belangrijke activiteit naast het onderzoek was GeNeYouS, het Genomics Network for Young Scientists. Gedurende de eerste drie jaar vanaf de oprichting heb ik me ingezet voor jullie. De bestuursvergaderingen, de ALV's, en alle georganiseerde activiteiten: het heeft me allemaal erg veel plezier opgeleverd. Ik wil dan ook alle mensen met wie ik in het bestuur heb gezeten bedanken voor de fijne periode: David, Terry, Simon, Mirre, Ilona, Roos, Wilbert, Fina, Christine, Maurice: het ga jullie goed. Ook de Internet/PR commissie wil ik hier noemen: Bernd, Rob, Martijn, Cordny, Hosea: bedankt voor de prettige samenwerking, ook al was deze vooral per e-mail (zoals het een commissie Internet/PR betaamt natuurlijk). Terry, de autorit naar Lyon, en het verblijf daar, was zeer geslaagd. Dat werd mede mogelijk gemaakt door je goede Frans, ook al was het met een Parijs' accent.

Een andere nevenactiviteit werd verricht in de NCMLS PhD committee. Hierin heb ik vooral veel geleerd van het organiseren van de jaarlijkse retraite, het eerste jaar met Els en het tweede jaar met Kirsten. Allebei bedankt voor de prettige samenwerking, evenals natuurlijk de andere leden van de commissie.

Waar ik misschien nog wel het meeste plezier uithaalde, was mijn grootste hobby: voetbal, en dan met name de Nijmegen Eendracht Combinatie, kortweg N.E.C.. Mijn passie voor deze vereniging kwam voornamelijk tot uiting in het bezoeken van alle thuiswedstrijden, de eerste jaren met Simon en later met broer Jeroen. 'Koetje' Simon, geweldig dat we samen op de Hazenkamp-tribune de voor- en nadelen van de wetenschap konden bespreken. Ook prachtig waren de bezochte uitwedstrijden bij 'jouw' SC Heerenveen en in de winderige 'fietsenstalling' van RKC Waalwijk, wedstrijden die nog (meestal) werden gewonnen ook. Via deze weg wil ik meteen de voltallige N.E.C.-selectie bedanken voor de mooie jaren, en dan met name voor het halen van

Europees voetbal in 2003. En wie ik ook niet mag vergeten: alle N.E.C.-webmasters. Laten we die jaarlijkse bowlingavond nog maar een lange tijd in stand houden!

Nog een aantal belangrijke mensen die ik nog niet genoemd heb:

Anand: tijdens mijn promotieperiode heb je mij geholpen met je vriendschap, en natuurlijk ook met de onvergetelijke rondreis door India, die je voor ons had geregeld via je familie en vrienden. Ook bedankt dus aan hen, en dan vooral aan je broers Swapnil en Chandan, je moeder Archana en je vader Kamalnayan. Ook geweldig dat je samen met David, Ruud en mij Indigonet gaat opstarten. Wat betreft die afgebroken vierdaagse: die lopen we nog wel een keer uit.

Chris, je feestjes op de Lange Hezelstraat waren altijd erg gezellig. Ook je bijdrage aan het wekelijkse uurtje voetbal in Park Brakkenstein mag niet worden vergeten. Jammer dat je nogal eens was verhinderd door een of andere vage blessure. Maar toch vooral bedankt dat ik bij je terecht kon op de moeilijke momenten van de afgelopen vier jaar. Ik wens je veel geluk toe in je nieuwe thuisland Litouwen.

Ana, alias Party Bee, je ‘workpaces’ vormden een zeer welkome afwisseling op het werk. Ook bedankt voor onze groepsvakantie in Valencia en omstreken. Losgelaten uit je ‘bubble’ bij Organon ga je nu zelf promoveren in Wageningen. Paella, cerveza, chica, adios!

Raoul, alias Berry, leuk met je te hebben gewerkt. Bij het Organon-dinsdagmiddag-zaalvoetbalteam heb je duidelijk gemaakt dat je niet voor niets de broer bent van een ex-proefvoetballer. Ik zal aan je denken als ik Jorien van den Herik of Sylvie Meijs op TV zie.

Greetje, alias Juffrouw Jannie, bedankt voor je immer goede humeur en diepgaande gesprekken. Zoals je ziet heb ik mijn promotie voltooid zonder enige ingewikkelde formule in het proefschrift. Bedankt voor de altijd lekkere koffie … en veel succes in de toekomst.

Nog wat mensen die ik al ken uit mijn studietijd maar nog steeds goede vrienden zijn: Rick, bedankt voor je aanwezigheid tijdens enkele verre reizen en Lowlands. Peter, bedankt voor de skivakantie in Tirol, ook al was het voor mij echt maar een éénmalig iets. Robert, leuk die pokeravonden en je kaartjes voor de skybox van FC Utrecht. Ook alle andere pokervrienden, bedankt voor alle pre-flop all-ins die me wat extra euro’s hebben opgeleverd om het promotiefeestje mee te kunnen financieren.

Nu we bijna op het einde zijn is het tijd voor mijn familie: vader Kees, moeder Corrie en broer Jeroen: allen bedankt voor jullie steun. Via dit boekje kan ik jullie nu eindelijk laten zien waar ik de afgelopen jaren mee bezig ben geweest.

Tenslotte, last but not least, natuurlijk Miaomiao:

我亲爱的小耗子，谢谢你无条件的爱和支持。虽然我们相知在我博士的最后阶段，但我知道如果没有你，一切都可能变得更加困难。在接下来的日子里，我会以相同的支持回报你。感谢你成为我博士答辩的助手。我还希望感谢你的父母和他们在一年期间电话里的支持,相信下个月我能亲自感谢他们。

Tim.

Curriculum vitae

Tim Hulsen werd op 10 augustus 1979 geboren te Wijchen. Zijn gymnasium-opleiding volgde hij aan het Dominicus College te Nijmegen, waar hij in 1997 zijn eindexamen behaalde. Vervolgens startte hij aan de Radboud Universiteit Nijmegen, destijds Katholieke Universiteit Nijmegen, met zijn studie Biologie. Er werden twee stages doorlopen, de eerste bij de afdeling Moleculaire Dierfysiologie onder leiding van Prof. Dr. Gerard Martens. Het onderwerp van de stage was de generatie van GST-p24 fusie-eiwitten voor antisera productie en p24/COP bindingsstudies. Een scriptie werd geschreven over het cAMP response element binding protein (CREB) en zijn rol in het zenuwstelsel. Een tweede afstudeerstage werd doorlopen in de bioinformatica, bij het Center for Molecular and Biomolecular Informatics (CMBI) onder leiding van Prof. Dr. Gert Vriend. Hier werkte hij aan de verheldering van de structuur van G Protein-Coupled Receptors (GPCRs) met behulp van de structuur van (bacterio)rhodopsine. Na zijn afstuderen in de zomer van 2001 ging hij verder als bioinformaticus, en wel als tijdelijke kracht bij NV Organon in Oss, afdeling Molecular Design and Informatics (MDI). Hier verrichtte hij ondersteunend werk voor verscheidene bioinformatica projecten. In november 2002 werd een begin gemaakt met het in dit proefschrift beschreven promotie-onderzoek op het CMBI, in een nauwe samenwerking met NV Organon. In het eerste jaar gebeurde dit onder leiding van Prof. Dr. Martijn Huynen, de overige drie jaar onder leiding van Dr. Peter Groenen. Per maart 2007 werkt Tim als post-doc in een Biorange project bij Radboud Universiteit Nijmegen / NV Organon.