OMA, A Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome Data: Introduction and First Achievements

Christophe Dessimoz*, Gina Cannarozzi, Manuel Gil, Daniel Margadant, Alexander Roth, Adrian Schneider, and Gaston H. Gonnet

ETH Zurich, Institute of Computational Science, CH-8092 Zürich
cdessimoz@inf.ethz.ch

Abstract. The OMA project is a large-scale effort to identify groups of orthologs from complete genome data, currently 150 species. The algorithm relies solely on protein sequence information and does not require any human supervision. It has several original features, in particular a verification step that detects paralogs and prevents them from being clustered together. Consistency checks and verification are performed throughout the process. The resulting groups, whenever a comparison could be made, are highly consistent both with EC assignments, and with assignments from the manually curated database HAMAP. A highly accurate set of orthologous sequences constitutes the basis for several other investigations, including phylogenetic analysis and protein classification.

Complete genomes give scientists a valuable resource to assign functions to sequences and to analyze their evolutionary history. These analyses rely heavily on gene comparison through sequence alignment algorithms that output the level of similarity, which gives an indication of homology. When homologous sequences are of interest, care must often be taken to distinguish between orthologous and paralogous proteins [1].

Both orthologs and paralogs come from the same ancestral sequence, and therefore are homologous, but they differ in the way they arise: paralogous sequences are the product of gene duplication, while orthologous sequences are the product of speciation. Practically, the distinction is very useful, because as opposed to paralogs, orthologs often carry the same function, in different organisms. As Eugene Koonin states it [2], whenever we speak of "the same gene in different species", we actually mean orthologs.

1 Previous Large-Scale Efforts

The systematic identification of orthologous sequences is an important problem that several other projects have addressed so far. Among them, the COG

* Corresponding author.

A. McLysaght et al. (Eds.): RECOMB 2005 Ws on Comparative Genomics, LNBI 3678, pp. 61–72, 2005.

database [3], [4] is probably the most established. From BLAST alignments [5] between all proteins ("all-against-all"), they identify genome-specific best hits, then group members that form triangles of best hits. Finally, the results are reviewed and corrected manually.

A further initiative is KEGG Orthology (KO) [6], [7]. KEGG is best known for its detailed database on metabolic pathways, but as the project evolved, an effort to cluster proteins into orthologous groups was initiated as well. The method is somewhat similar to COG: it starts with Smith-Waterman [8] all-against-all alignments, and identifies symmetrical best hits. It then uses a quasi-clique algorithm to generate "Ortholog clusters", that are used to create the KO groups, the last step being performed manually.

Finally, we mention here Inparanoid [9], OrthoMCL [10] and EGO (previously called TOGA) [11]. All three projects exclusively cover eukaryotic genomes. The two first insist on the inclusion of so-called "in-paralogs", sequences that result from a duplication event that occurred after all speciations. A noticeable shortcoming of Inparanoid is the fact that it only handles pairs of genomes at a time. As for EGO, although their last release contains almost half a million genes from 82 eukaryotes, many sequences appear in more than one group and many groups contain paralogs. Because of that, we consider Inparanoid and EGO outside the present scope and limit our comparisons below to COG, KO and OrthoMCL.

2 Overview of the OMA Project

The project presented in this article is a new approach to identify groups of orthologs. It has some very specific properties:

- *Automated.* Unlike COG and KEGG Orthology, the whole workflow does not require human intervention, thereby insuring consistency, scalability and full transparency of the process.
- *Extensive.* The analysis so far has been performed on more than 150 genomes (Prokaryotes and Eukaryotes), with new ones added by the day[1]. The goal is to include all available complete genomes.
- *Strict.* Consistency checks are performed throughout the workflow, particularly at the integration step of genomic data. The algorithm for the identification of orthologous proteins excludes paralogs. 98.3% of the groups we could test are made of *bona fide* orthologous proteins (Sect. 4.1).

The algorithm for the identification of orthologous groups relies solely on protein sequence alignments from complete genomes, and hence does *not* depend on previous knowledge in terms of phylogeny, synteny information or experimental data. It is described in detail in the next section.

From the orthologous groups, we build a two-dimensional matrix in which each row represents an orthologous group and each column represents a species.

[1] At the time the final version of this article is submitted, 181 genomes have been included in the analysis.

The applications of that matrix are numerous and fall beyond the scope of this article. However, a few are worth mentioning. The rows provide phyletic patterns of the orthologous groups and can be used for phylogenetic profiling [12]. Parsimony trees can be constructed from the matrix to give either a phylogenetic tree when built from the columns, or protein families when built from the rows. We believe that both trees are very valuable contributions, and they will be presented, among others, in separate articles. Also, a large set of orthologous sequences is a prerequisite for the construction of reliable phylogenetic distance trees.

3 Methods

The construction of the matrix is performed in four steps. In the first one, genomic data is retrieved, checked for consistency and integrated. The second step consists of Smith-Waterman [8] protein alignments between all proteins ("all-against-all") followed by the identification of stable pairs, essentially what is sometimes also referred to as "symmetrical best hits". In the third step, the algorithm verifies every stable pair to ensure that it represents an orthologous relationship, not a paralogous one. Finally, in the fourth step, groups of orthologous proteins are formed from cliques of verified stable pairs.

3.1 Genome Data Retrieval, Verification and Integration

Complete genomes with protein sequence information are retrieved from Ensemble [13] and GenBank [14] and checked for consistency, then imported into *Darwin* [15], our framework. The consistency verification is extensive, and includes comparison between DNA and amino acid sequence, check for presence of start and stop codon, removal of fragments shorter than 50 amino acids, removal of duplicated sequences (sequences with >99% identity), verification of the total number of entries with HAMAP [16] (or GenBank/Ensembl for eukaryotes), and comparison with sequences present in SwissProt [17]. In case of alternative splicing, only the largest set of non-overlapping splice variants is kept for further analysis.

3.2 All-Against-All

Every protein sequence is aligned pairwise with every other protein sequence from a different organism using full dynamic programming [8]. The alignments were performed with GCB PAM matrices [18], using, for each alignment above noise level, the matrix corresponding to the PAM distance that maximizes the score, in a maximum likelihood fashion [19]. Alignments with score below 198 (70 bits, which typically corresponds to an E-value around 1.3e-16) or with length below 60% of the smaller sequence are considered not significant, and are discarded. The use of BLAST [5] was evaluated, but in the present case, we considered the speed increase not sufficient to compensate the loss in sensitivity [20]. Note that this view is shared by the teams behind KEGG Orthology [7] and STRING [21].

From the alignments, stable pairs are identified. That is essentially the idea behind what COG, among others, call "symmetrical best hits", that is, a protein pair in two different organisms that have each other as best match. However, as opposed to them, we improve robustness by keeping matches that have scores not significantly lower than the best match. Concretely, a stable pair can be formed between two proteins in two different organisms if, in both directions, the score of the alignment is not less than 90% of the best match.

3.3 Stable Pairs Verification

At this point, most stable pairs are expected to link two orthologous proteins, because orthologs usually have a higher level of similarity than paralogs. However, in case the corresponding ortholog of a particular protein is missing in some species (e.g. the organism lost it during evolution), a stable pair might be formed between that protein and a paralogous sequence, thus linking two proteins that belong to different orthologous groups. Such instances can be detected through the comparison to a third species that carries orthologs to both proteins (Figs. 1, 2). Therefore, each stable pair is verified through an exhaustive search against every other genome for such a scenario, and stable pairs corresponding to paralogy are discarded (Fig. 4). A more formal description of this algorithm, with proofs and examples are part of a separate publication.

3.4 Group Construction from Cliques of Verified Stable Pairs

The last step consists of orthologous groups identification from all verified stable pairs. The problem can be seen as a graph where proteins are represented by vertices and stable pairs by edges. In such a graph, an orthologous group is expected to form a fully connected subgraph. Thus, the algorithm iteratively looks for the maximal clique, groups the corresponding proteins and removes them from the graph. It runs until no more verified stable pairs are left. Finding

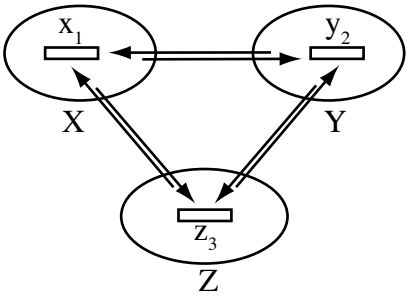

Fig. 1. Orthologous relationship between x_1 and y_2

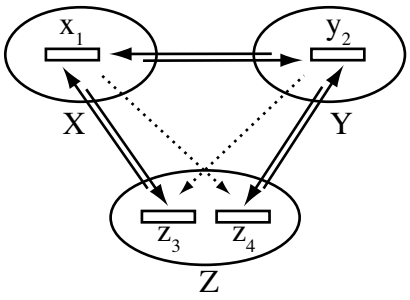

Fig. 2. Paralogous relationship between x_1 and y_2, demonstrated by presence of z_3 and z_4 in Z

maximal cliques is a difficult problem (NP complete). The implementation of clique finding in *Darwin* [15] is based on the vertex cover problem and is a very effective clique approximation, which runs in reasonable time [22].

3.5 Tests for Accuracy and Completeness

On a project of such large size, it is crucial to ensure that all steps have been performed correctly, and that nothing is missing. With more than a hundred computers working around the clock for months, the probability of technical and operational failures becomes non-negligible, and must be proactively managed. We have included a number of tests that ensure quality all along the procedure described above. One test verifies that alignments are not missing through random sampling of 50,000 alignments per pair of genomes. Another test completely recomputes all recorded alignments of a pair of genomes, which is useful to detect (rare) errors due to hardware failure. A signature of the genomic database is computed at the end of each run to insure that memory was not corrupted during the computation. Yet another test verifies consistency of the results by looking for triangles of stable pairs that have a missing edge. More than once, these tests have revealed missing data, faulty hardware, and bugs in our programs.

4 Results and Discussion

The last OMA release classifies 501,636 proteins from 150 genomes into 111,574 orthologous groups (called *OMA groups* below). That covers 65.81% of all proteins contained in those genomes. The distribution of group size is such that most groups are small (Fig. 3). To a large extent, that is an obvious consequence of the large biodiversity among the included genomes. However, a technical reason can also explain part of that phenomenon: relatively few higher eukaryotes, and in particular plants, have been sequenced and included at this point, but they represent a significant portion of the total genes. All plant-specific genes in the matrix currently belong to groups of size two, simply because only two plants (*Arabidopsis thaliana* and *Oryza sativa*) are present. This effect is also reflected by a lower coverage of some eukaryotes. Therefore, we expect the group size and coverage to increase as more genomes are included.

The average group size is compared to other projects in Table 1. The differences are considerable. They can be explained by at least four factors: i) Quality of the algorithm. ii) Difference in the treatment of paralogous sequences. COG, KO, HAMAP and OrthoMCL often classify more than one protein per species into the same group. These proteins cannot have an orthologous relationship, by definition. In the best cases, those proteins are in-paralogs, genes that result from a duplication after all speciations, where justification for such inclusion is usually that in-paralogs are orthologous to all other proteins in the group. iii) Human validation. The practical problems of managing many groups are likely to create a bias toward fewer, larger groups (that can be observed in Table 1).

iv) Variation in the species composition and more generally in the biodiversity of the included sequences.

In that context, the size of the groups assigned by our algorithm does not appear unreasonable.

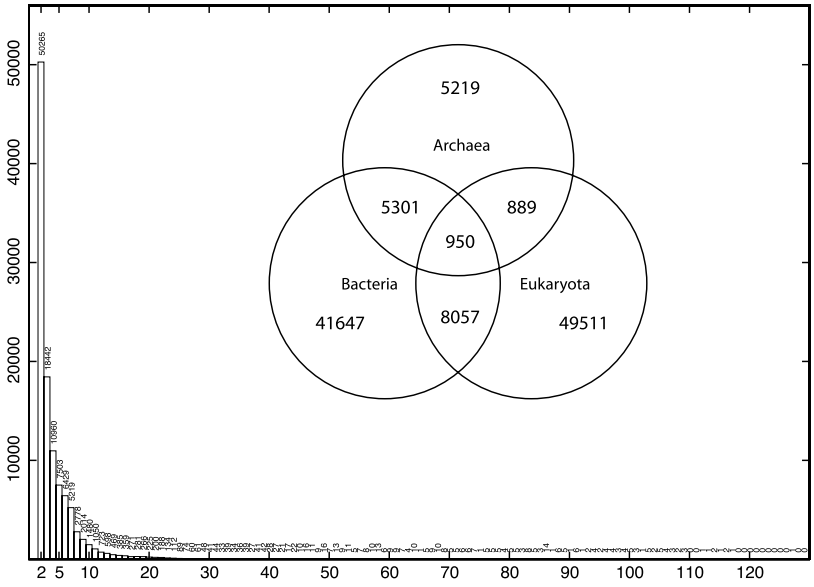

Fig. 3. Histogram of orthologous groups size and repartition of the 111,574 groups among kingdoms

Table 1. Comparison of some statistics accross projects. Note that KO and HAMAP only include partial genomes.

Project Name	Release	#Species	#Seqs	#Groups	Average Group Size	Coverage
COG	2003	66	138,458	4,873	28.4	75%
KO	22/Apr/2005	244	284,519	5,795	49.1	n/a
HAMAP	30/Apr/2005	876	26,977	1,071	25.2	n/a
OrthoMCL	I=1.5, 2003	7	47,668	7,265	6.6	47%
OMA	13/May/2005	150	501,636	111,574	4.5	66%

4.1 Validation

The quality of the groups resulting from our algorithm must be ensured. The statistics above about group size and genomes coverage constitute a first check,

but more specific analysis of the results are desirable. This section presents the results from two further verifications, one using Enzyme Classification nomenclature, the other comparing our results with manual ortholog assignments from expert curators.

Function Validation Using Enzyme Classification. Enzyme Classification (EC) numbers are assigned based on the enzymatic activity of proteins. Since orthologs usually keep the same function, we expect in general that enzymes belonging to the same OMA group all have identical EC number. The Swiss Institute of Bioinformatics maintains the database [23] on Enzyme nomenclature that served us as reference (Release 37.0 of March 2005). First, the proteins that have more than one EC number (multi-functional enzymes, about 3% of all sequences in the EC database) were removed from the analysis. Then, every OMA group with at least two proteins that could be mapped to the EC database were selected for comparison.

There were 2,825 such groups out of 111,574 groups (2.5%). Of those, 2750 groups (97.3%) mapped to a single EC class. That compares very favorably to OrthoMCL, that has only 86% of its groups consistent with the EC assignments [10], although in their analysis, multi-functional enzymes were not excluded[2]. The result obtained for our method is particularly good if we consider that not all orthologs have identical function [24], and that the EC database is most probably not completely error-free.

Table 2. Comparison with HAMAP families

OMA Groups corresponding to HAMAP families:	1993	100%
— mapping to a single family:	1959	98.3%
— mapping to more than one family:	34	1.7%
HAMAP families corresponding to OMA Groups:	974	100%
— mapping to a single group:	355	36.4%
— mapping to more than one group:	619	63.6%

Comparison with HAMAP. Our groups were also compared with those of the HAMAP project [16]. As stated on their website, the HAMAP families are a collection of orthologous microbial proteins generated manually by expert curators. The comparison was done as following: in each HAMAP family, the inparalogs were removed. OMA groups that had at least two proteins linkable to HAMAP were considered. Conversely, the HAMAP families with at least two proteins linkable to OMA groups were kept. Then, the correspondence between both sets of groups was assessed (Table 2). The results clearly show that while

[2] To compare the results with OrthoMCL in all fairness, the same analysis was performed on an OMA release from 26 eukaryotes, without removing multi-functional enzymes. With 1054 out of 1082 groups (97.4%) mapping to a single EC class, there was practically no difference.

our algorithm generates more/smaller groups than HAMAP, these groups are almost always consistent with the HAMAP assignments, in the sense that they each map to a single family. In fact, the numbers are such that slightly more than a third of the HAMAP families (347 out of 974) have a 1:1 correspondence to our groups, while most of the other two thirds are covered by typically two or three OMA groups. Considering that HAMAP families and OMA groups are constructed using radically different methodologies, this level of consistency is remarkable.

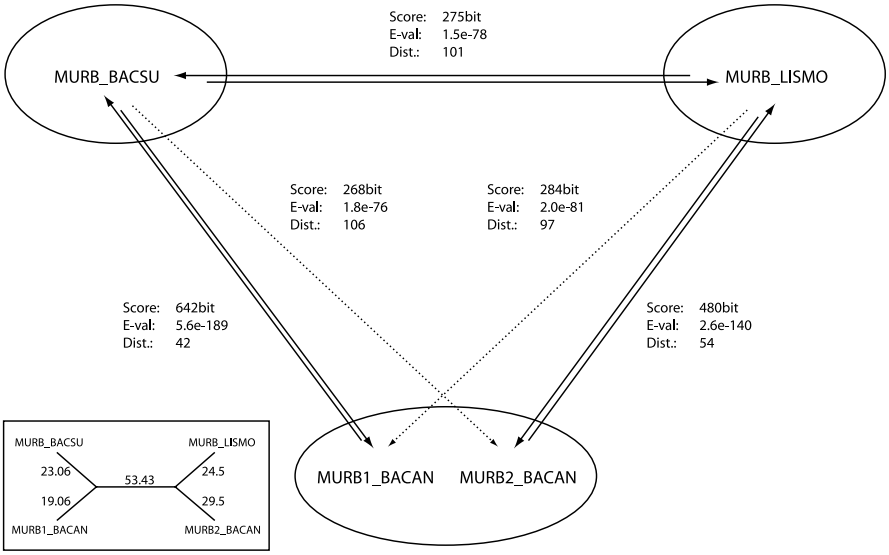

Fig. 4. Paralogs inside the HAMAP family MF_00037 (distances in PAM units)

The question that naturally arises from the comparison is whether it is our algorithm that has an excessive tendency to split orthologous groups or it is HAMAP that forms too large families. We performed some case-by-case analysis that revealed dubious classification on both sides: we have found several instances of OMA groups that have been split as a result of missing stable pairs (typically caused by alignment scores or length below our current threshold). Conversely, we found instances of sequences very likely to be paralogous within the same HAMAP family (Fig. 4). At this point, we are still investigating the relative merits of tighter versus larger groups.

4.2 Computational Cost

The all-against-all is the most time-consuming part of the computation. From the 150 genomes, we have 762,265 proteins producing about 2.85e11 pairwise alignments. In terms of the dynamic programming algorithm, the number of

cells is 4.16e16 (where a cell corresponds to the computation of the table entry in the dynamic programming algorithm to align two sequences). We use *Darwin* [15] in parallel on more than 200 CPUs, which gives us a total capacity of about 1.9e13 cells/h (a Pentium IV at 3.2 GHz can perform about 2.2e10 cells/h). Hence, under such conditions, about 91 days would be required to compute the all-against-all. In practice, it took us longer, because of the consistency checks and changes in the data or programs. As for the stable pair verification and clique algorithm, they can be computed in about two days on a single machine.

4.3 Availability of the Results

The project homepage can be reached under `http://www.cbrg.ethz.ch/oma`. The list of species, progress of the all-against-all and OMA groups statistics are updated continuously. We offer a prototype online interface that enables users to browse through the results online.

5 Open Problems

As stated previously, the project is ongoing and some issues remain to be addressed. One of them concerns how to handle multi-domain proteins. The question is important, because the majority of proteins in Prokaryotes and Eukaryotes consist of at least two domains [25], where a domain is defined as an independent, evolutionary unit that can either form a single-domain protein or be part of a multi-domain one. Currently, our algorithm classifies a multi-domain protein with the group of its highest scoring domain. While that does not cause disruptive harm, it gives incomplete information about that multi-domain protein. In terms of consistency, it is not desirable to have that protein grouped with orthologs of its best scoring domain, while not grouped with orthologs of, say, its second best scoring domain. Either the focus is on domain orthology and it should be grouped to both, or the focus is on whole protein orthology and it should be grouped with none.

Lateral gene transfer is also a potential source of complications. Despite the abundant literature on the subject, the actual extent of this phenomenon remains unclear. Here as well, the effect on our group building process is non disruptive, xenologs are currently merely included in orthologous groups, but might cause problems in applications sensitive to phylogeny (e.g. phylogenetic trees). We are working on methods to systematically identify potential cases of lateral gene transfer *a posteriori*. The details and conclusions of this work will be the object of a separate publication.

6 Conclusion

The systematic identification of orthologuous sequences is an important problem in bioinformatics. In this article, we have presented OMA, a new large-scale project to cluster proteins into orthologous groups, where both the amount of

data (150 genomes) and amount of computation (>500,000 CPU hours) justifies the large-scale description. Strict verification and consistency checks are performed throughout the workflow. The orthologous group construction is performed by an algorithm with several original features: it estimates a PAM distance between pairs of sequences matching significantly, it extends the concept of symmetrical best hit by considering all possible pairs of top matches within a tolerance factor, it detects and discards stable pairs connecting paralogous sequences and finally it identifies cliques of stable pairs to construct the groups. In contrast to most other projects, it does not rely on human validation. The resulting groups are highly consistent with EC assignments whenever applicable. They are also highly consistent with the manually curated database HAMAP, although our algorithm seems to have a tendency to split orthologous groups excessively. That issue, along with handling of multi-domain proteins and detection of lateral gene transfer events are the main problems that remain unsolved for now. However, even in its present state, we are confident that the project is an important contribution toward better identification of orthologous groups, and that it constitutes a solid basis for future work.

Acknowledgements

The authors thank Jean-Daniel Dessimoz, Markus Friberg and three anonymous reviewers for their comments and suggestions on the present manuscript, as well as Brigitte Boeckmann and Tania Lima from the Swiss Institute of Bioinformatics for useful discussions.

References

1. Fitch, W.M.: Distinguishing homologous from analogous proteins. Syst Zool **19** (1970) 99–113
2. Koonin, E.V.: An apology for orthologs - or brave new memes. Genome Biol **2** (2001) COMMENT1005
3. Tatusov, R.L., Koonin, E.V., Lipman, D.J.: A genomic perspective on protein families. Science **278** (1997) 631–7
4. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A.: The cog database: an updated version includes eukaryotes. BMC Bioinformatics **4** (2003) http://www.biomedcentral.com/1471–2105/4/41
5. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25** (1997) 3389–3402
6. Fujibuchi, W., Ogata, H., Matsuda, H., Kanehisa, M.: Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. Nucleic Acids Res **28** (2000) 4029–4036

7. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M.: The KEGG resource for deciphering the genome. Nucleic Acids Res **32** (2004) 277–280
8. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. **147** (1981) 195–197
9. Remm, M., Storm, C., Sonnhammer, E.: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol **314** (2001) 1041–52
10. Li, L., Stoeckert, C.J.J., Roos, D.S.: OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13** (2003) 2178–2189
11. Lee, Y., Sultana, R., Pertea, G., Cho, J., Karamycheva, S., Tsai, J., Parvizi, B., Cheung, F., Antonescu, V., White, J., Holt, I., Liang, F., Quackenbush, J.: Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res **12** (2002) 493–502
12. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96** (1999) 4285–4288
13. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., Clamp, M.: The Ensembl genome database project. Nucleic Acids Res **30** (2002) 38–41
14. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L.: GenBank. Nucleic Acids Res **33 Database Issue** (2005) 34–38
15. Gonnet, G.H., Hallett, M.T., Korostensky, C., Bernardin, L.: Darwin v. 2.0 an interpreted computer language for the biosciences. Bioinformatics **16** (2000) 101–103
16. Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A.H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C.J.A., Lachaize, C., Veuthey, A.L., Gasteiger, E., Bairoch, A.: Automated annotation of microbial proteomes in SWISS-PROT. Comput Biol Chem **27** (2003) 49–58
17. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res **31** (2003) 365–370
18. Gonnet, G.H., Cohen, M.A., Benner, S.A.: Exhaustive matching of the entire protein sequence database. Science **256** (1992) 1443–1445
19. Gonnet, G.H.: A tutorial introduction to computational biochemistry using Darwin. Technical report, Informatik, ETH Zurich, Switzerland (1994)
20. Brenner, S.E., Chothia, C., Hubbard, J.T.: Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A **95** (1998) 6073–6078
21. von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., Bork, P.: STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res **33 Database Issue** (2005) 433–437
22. Balasubramanian, R., Fellows, M.R., Raman, V.: An improved fixed-parameter algorithm for vertex cover. Inf. Process. Lett. **65** (1998) 163–168
23. Bairoch, A.: The ENZYME database in 2000. Nucleic Acids Res **28** (2000) 304–305

24. Jensen, R.A.: Orthologs and paralogs - we need to get it right. Genome Biol **2** (2001) INTERACTIONS1002
25. Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., Teichmann, S.A.: Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol **14** (2004) 208–216