ELSEVIER

Computers and Chemistry 000 (2001) 1–7

Computers & Chemistry

www.elsevier.com/locate/compchem

Medical target prediction from genome sequence: combining different sequence analysis algorithms with expert knowledge and input from artificial intelligence approaches

Thomas Dandekar [a,b,c,*], Fuli Du [a,b], R. Heiner Schirmer [d], Steffen Schmidt [a,b]

[a] *European Molecular Biology Laboratory, PO Box 102209, Meyerhostraße 1, D-69012 Heidelberg, Germany*
[b] *Parasitology Center, INF, 69120 University of Heidelberg, Heidelberg, Germany*
[c] *Institute for Molecular Medicine, 79106 University of Freiburg, Freiburg, Germany*
[d] *Biochemistry Center (BZH), INF, 69120 University of Heidelberg, Heidelberg, Germany*

Abstract

By exploiting the rapid increase in available sequence data, the definition of medically relevant protein targets has been improved by a combination of: (i) differential genome analysis (target list); and (ii) analysis of individual proteins (target analysis). Fast sequence comparisons, data mining, and genetic algorithms further promote these procedures. *Mycobacterium tuberculosis* proteins were chosen as applied examples. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Genome analysis; Structure prediction; *Mycobacterium tuberculosis*; Drug design; Sequence comparison

1. Introduction

Revealing protein targets from genome screening is a new major challenge (Emilien et al., 2000). For individual molecules or even a family of related structures, efficient procedures incorporating methods from artificial intelligence exist (e.g. Ghuloum et al., 1999; Polanski and Walczak, 2000 for recent QSAR studies). Nevertheless, in areas such as protein structure prediction, a combination of expert knowledge with computer automation is often used (Bates et al., 1997; Trohalaki et al., 2000). In genomes, the combination of specific databases, sequence search algorithms and expert knowledge offers a promising approach to tackle the higher complexity of a genome compared to a single protein (Koonin, 2001). Our approach presented here is medically oriented. It combines different sequence analysis algorithms with expert knowledge to divide and analyze the complete set of proteins in a genome regarding pharmacologically motivated sets and target categories. Simple routines from artificial intelligence approaches are included to enhance this complex prediction task. Next, individual targets are efficiently analyzed including model building, homology modelling and domain retrieval. This is demonstrated for proteins and structures from different target categories in *Mycobacterium tuberculosis*. Presently, our combined approach allows not only efficient target retrieval for pharmacologically interesting protein categories (individual examples are presented), but also a substantial gain in genomic protein function assignment compared to direct annotation (about one-third less unassigned proteins). This is compared to other recent advances in computational genomics. Further gains from context-based and artificial intelligence-based approaches are expected (Tsoka and Ouzounis, 2000).

* Corresponding author. Tel.: + 49-6221-387-466; fax: + 49-6221-387-306.
E-mail address: dandekar@embl-heidelberg.de (T. Dandekar).

0097-8485/01/$ - see front matter © 2001 Elsevier Science Ltd. All rights reserved.
PII: S0097-8485(01)00095-X

2. Materials and methods

2.1. Differential genome analysis

Differential genome analysis was carried out extensively using sequence comparison methods as described (Huynen et al., 1998; Dandekar et al., 2000). Briefly, lists comparing similarities were created by sensitive sequence alignment algorithms for the proteins from the complete genome sequence. Relatedness was established by significant expected values ($E > 0.001$) in sequence comparisons. It was carefully checked whether this applied to the whole sequence or only parts of the complete protein. Separate protein domains with different predicted function were distinguished by detailed sequence analysis. Orthologs were determined by comparing complete genomes from different organisms. Homologous sequences are called orthologous when their independent evolution reflects a speciation event rather than a gene duplication event (Fitch, 1970). Orthologous sequences are likely to have the same function in the two species compared. In the determination of orthology, we use both relative similarity (two orthologous sequences should have the highest pairwise identity level, compared to the level of identity of either of the sequences to all other sequences in the other sequence's genome), and relative position within the genome (conservation of gene neighborhood). Several sequence analysis algorithms were used for this (BLAST, FASTA and PSI-BLAST; Aravind and Koonin, 1999). In particular, with application-specific hardware for large-scale sequence to sequence comparisons, the speedup obtained is large enough to compare even complete prokaryotic genomes such as different *Mollicutes* among each other (Dandekar et al., 2000).

2.2. Secondary structure prediction

Profile-based neural networks were used applying Rost's program PHD (Rost and Sander, 2000). The program PREDATOR (Persson, 2000; Frishman and Argos, 1997) served to achieve a secondary structure prediction based on pairwise alignment of the sequence to be predicted with each related sequence. Secondary structure assignment comparing the blosum62 similarity scores (Pearson, 2000) of best-matching fragments in a database of known structures was done using the program SIMPA 96 (Levin, 1997). Secondary structure predictions were compared and combined for model building in the case of a target structure where no tertiary template was known. Model coordinates are available on request.

2.3. Homology modelling and domain identification

Homology modelling used the PEITSCH software package available from SwissModel (Schwede et al., 2000). Domain identification applied iterative sensitive sequence alignment procedures including PSI-BLAST (Aravind and Koonin, 1999) and the simple modular architecture detection tool (SMART; Schultz et al., 2000) and followed detailed protocols as described previously (Dandekar et al., 2000; Bork et al., 1998). Identified known three-dimensional (3-D) structures were classified according to SCOP (Lo Conte et al., 2000).

2.4. Optimization of different ranking criteria using the genetic algorithm

Since ranking and evaluation of drug or target candidates can be difficult because of the great number of criteria and compounds to be considered, we wrote a genetic algorithm optimizer for this task.

Input file: Different pharmacologically interesting compounds (or drug targets) are ranked according to their established quality on the basis of a gold standard for the desired pharmacological activity, for instance, their antibiotic efficacy. In addition, they are measured in their quality (simple scalar values such as point scores or floating point values) by different physico-chemical criteria such as hydrophobicity, antigenicity, content of secondary structure and accessibility. These data are stored in a flat file table.

Our program utilizes a genetic algorithm (Pena-Reyes and Sipper, 2000)-driven search engine to optimize the weights for the different criteria. The program was written in PASCAL. Weights were initially randomly chosen and later optimized and selected on comparing arrays with different target ranking according to the encoded weights. They were evolved in successive generation cycles in such a way that the weights chosen for the different other criteria can best predict the gold standard ranking (e.g. antibiotic efficacy).

Running conditions: After a random start, high quality bit-strings are selected preferentially as parents. They are mutated using on average one bit per string per generation and recombined through crossover; the probability of recombination is 0.2 per bit string per generation and occurs at exactly one equivalent site chosen at random on each of the parental chromosome pairs. This yields the next parental generation of encoded weights. A positive constant keeps the population of prediction trials richer since low-fitness individuals may also survive. Simulations were run over 300 generations to allow convergence.

3. Results and discussion

3.1. Differential genome analysis

Large-scale sequencing projects now yield a huge amount of data including the complete genome sequences of a number of prokaryotic and eukaryotic organisms. Here, we will concentrate on the severely pathogenic genome from *M. tuberculosis*.

Detection of open reading frames is, even in prokaryotes, a non-trivial exercise. Software such as GENESCAN (Ramakrishna and Srinivasan, 1999) is used; recent advances involve data mining for more sensitive detection (King et al., 2000).

The next step for protein target identification is differential genome analysis. It tackles the following recognition task: which proteins are specific for this organism, which are shared with several other species and which proteins occur ubiquitously? In particular, genome-specific enzymes may be pharmacologically targeted without hurting the human patient. Furthermore, genes shared only among pathogens often give a first clue for identifying new pathogenicity factors.

These questions can be answered by a two-step procedure; first, the group of genes or even a whole genome is compared by fast automated sequence comparison procedures with the corresponding genes from other genomes.

Three different types of genes can be distinguished. Genes from other species can be either (i) highly similar over most of the sequence and probably encode a protein with the same function, an ortholog; (ii) related only in some part of the complete sequence; or (iii) not significantly related.

Next, the genes encoding orthologous (see M&M) proteins, encoding probably the same function, are classified according to Venn diagrams. Different categories are identified in this way such as organism-specific genes (set 1, e.g. an organism-specific kinase), genes shared between several pathogenic species (set 2, e.g. host interaction factors), between most bacteria (set 3, e.g. ribosomal proteins) or between patient and parasite (e.g. triosephosphate isomerase). This Venn classification can be automated using awk scripts and PERL programs.

3.2. Lists of potential targets and further identification tools

Currently, we investigate improved ways to cluster and sort genes. Data mining subroutines look in addi-

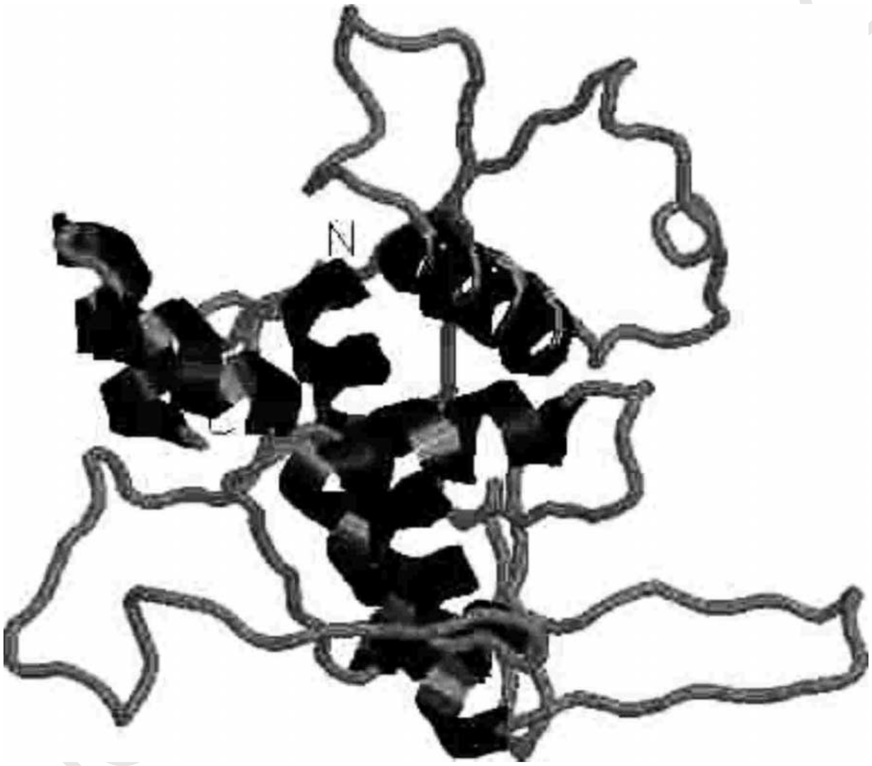

Fig. 1. Cartoon representation (using the program RASMOL from Sayle, see latest update by Bernstein (2000)) of the main chain trace of a deduced first approximation model of polyphosphate glucokinase from *M. tuberculosis*. N- and C- terminals are labeled. Helices are shown in black.

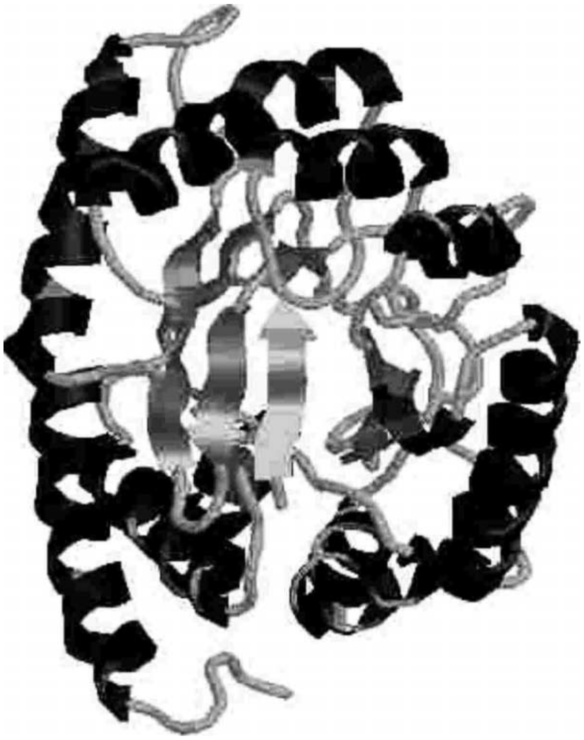

Fig. 2. Cartoon representation (using the program RASMOL, Bernstein, 2000) of the 3-D structure of the known SCOP domain 3.1.21.1.1 evident in the predicted dihydropteroate synthase homolog from *M. tuberculosis*. Helices are shown in black, sheet regions in dark grey.

tion to direct sequence similarity (orthologous genes, detected by the programs mentioned above) for similar functionality. This can be achieved by combining clusters according to sequence similarity with PERL programs that link and sort genes by functional patterns (e.g. classification, substrates) identified in the description line comparing genes from different species.

Moreover, simple rules such as the percentage of sequence identity multiplied by the length of the identity stretch are combined with enzyme-specific data strings (such as 'kinase') and recognition motifs (e.g. from the database PROSITE, Apweiler et al., 2001). This allows a far more specific recognition and classification of related enzyme activities with the potential for large scaleup. Thus, *M. tuberculosis* polyphosphate glucokinase (E.C. 2.7.1.63) can be readily identified and grouped into the context of other, not functionally identical but related enzyme activities, for instance, hexokinases and glucokinases from different organisms. Coupling sensitive sequence comparisons to straightforward text mining routines has been shown to significantly enhance structure prediction for remote homologs by MacCallum et al. (2000) (program SAWTED). Our combination of such routines together with latest software, genome context, detailed sequence analysis and expert knowledge to achieve protein function assignments allowed a substantial gain, for example, in *Mycoplasma pneumoniae* (Dandekar et al., 2000): 688 protein frames are identified and only 230 or 33% have no function assignment. Standard techniques (Himmelreich et al., 1996) identified only 677 protein reading frames and 328 or 48% remained with unknown function.

In particular, signal to noise ratios for protein function detection are improved and misannotations are reduced. The combination of genome context with different algorithms and databases is the main reason for this gain (compare with similar results reported by Koonin (2001), Kyrpides et al. (2000)). Regarding artificial intelligence techniques, one has to stress that the routines applied are still simple and human expert knowledge and intervention includes critical steps. The gain nevertheless achieved indicates that further developments in this direction should allow an even larger improvement in sequence assignment as well as in prediction and analysis of valuable targets (see below). This includes incorporation of techniques from data mining and symbolic machine learning (Helma et al., 2000) or inductive logic programming (to achieve probabilistic prediction rules; King et al., 2001).

3.3. Target structure analysis

Analysis of individual targets suggested from our genome analysis of *M. tuberculosis* is illustrated by structures obtained from the different protein sets.

The gene for polyphosphate glucokinase and its encoded enzyme are a specific adaptation of *M. tuberculosis* and not present in humans (see set 1 above, species-specific adaptations). This enzyme catalyzes the reaction: phosphate$_{(n)}$ + D-glucose ↔ phosphate$_{(n-1)}$ + D-glucose phosphate and presents a potential pharmacological target. Polyphosphate glucokinase from *M. tuberculosis* is not homologous to any known 3-D structure. We tested this using different sensitive sequence alignment programs (Aravind and Koonin, 1999) as well as by looking for structurally related protein domains with the SMART program (Schultz et al., 2000). However, a first approximation of the 3-D structure of polyphosphate glucokinase is helpful for drug design. Initial results applying secondary structure prediction, analysis of loop regions, accessibility (according to PHD; Rost and Sander, 2000) and selection of predictions from model building by different target evaluation criteria (compactness, protein topology; comparison to non-homologous enzymes of carbohydrate metabolism) suggest the globular fold we show in Fig. 1. Such models have to be improved in new structure prediction cycles (e.g. by applying genetic algorithms, by model building, or energy refinement) according to the experimentally obtained data and available biological knowledge (Saxena et al., 2001; Ota et al., 1999). After this, they can achieve reliable topology predictions (handedness, correct number and succession of secondary structure elements, etc.) and sufficiently low RMSD (around 6 Å and less; compare to recent improvements in pure ab initio prediction by Pillardy et al. (2001)).

The dihydropteroate synthase homolog in *M. tuberculosis* is an example for set 3 (general proteins) as it is present in many bacteria. However, the enzyme does not occur in humans. A comparison among more organisms including humans would put the protein away from the central set of proteins shared among all compared species. It provides the second step in dihydrofolate synthesis. The homolog is directly identifiable by sensitive sequence and domain comparisons from the raw genome sequence; see also the annotation from the *M. tuberculosis* sequencing consortium (Cole et al., 1998). Further, by straightforward domain analysis the protein can be shown to contain a SCOP domain of a known structure shown in Fig. 2. This enables, in this simple illustration example, a more accurate examination of the 3-D structure of the fold (after refinement RMSD error estimated to be 2–3 Å) to better fight the sulfonamide resistance of *M. tuberculosis*.

Pyruvate kinase from *M. tuberculosis* is an example for set 2 (proteins found in some species, but a clear subset of all compared) Fig. 3. A homology model is easily obtained (many related sequences and a structural homolog with detailed 3-D structure is available), e.g. using the Swiss modelling server (Schwede et al., 2000; RMSD error estimated to be 1–2 Å). The model shows minor differences when compared to other pyruvate kinases. In contrast to the above example, this enzyme occurs in humans and thus much more care is needed to exploit the predicted structure as a potential therapeutical target.

For the evaluation of different targets or drugs for a specific target by different criteria, many different parameters and rankings are possible. We developed a genetic algorithm-based optimizer for this. According to the desired criteria and a 'gold standard', e.g. antibiotic activity against a pathogen, optimal weights yield-

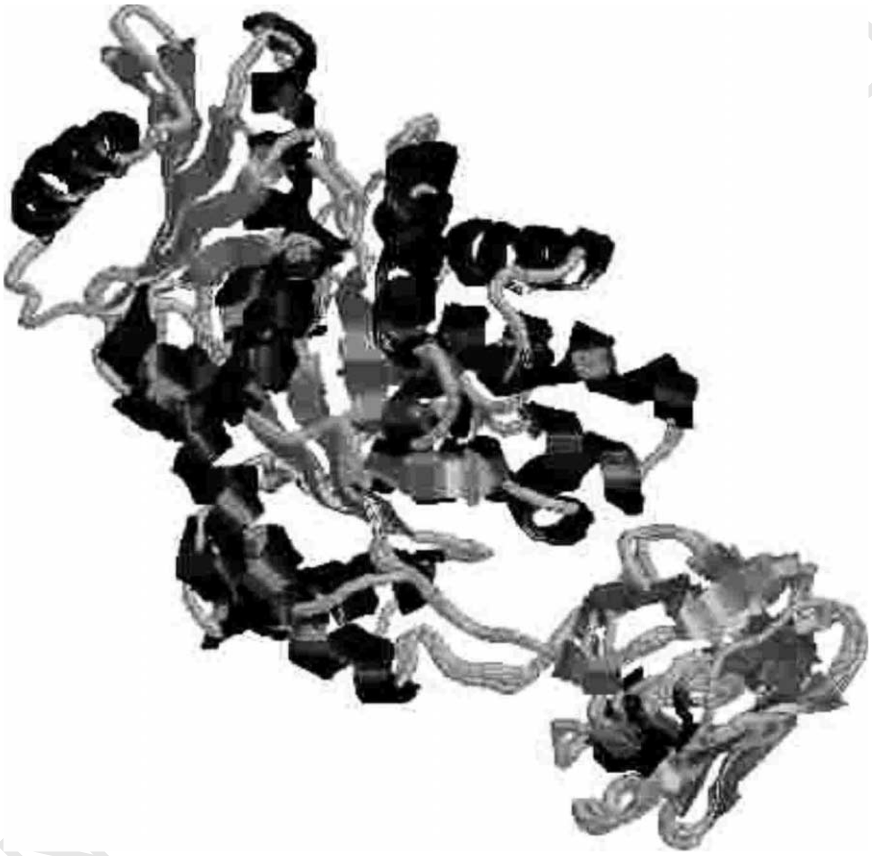

Fig. 3. Cartoon representation (using the program RASMOL, Bernstein, 2000) of a homology model of pyruvate kinase from *M. tuberculosis*. Helices are shown in black, sheet regions in dark grey.

Table 1
Target ranking illustration example

Overall hydrophobicity	N-terminal negative charge	Molecular weight	Relative toxicity	Relative antibiotic strength (average number of bacterial colonies remaining after antibiotic)
−16897.13	−6769.86	1979.00	3	9.8
−16650.64	−6676.40	1979.00	3	9.3
−15340.80	−5767.24	1852.00	4	11.3
−15339.99	−5765.54	1852.00	4	11.3
−15297.65	−5786.51	1940.00	5	8.2
−15398.68	−5803.39	1955.00	4	12.1
−15905.19	−5824.71	1965.00	4	7.0
−15252.84	−5641.23	1960.00	5	10.0
−17024.36	−6258.18	2009.00	0	6.6
−15272.44	−5921.17	1908.00	0	6.9
−14292.27	−5938.48	1998.00	4	7.1

…(further data) …

Optimized weights for predicting antibiotic efficacy of further compounds from the library using the other parameters after weight normalization to the first parameter:

1.000	11.69	3.35	−10.21

Many slightly different antibiotic compounds from a substance library are tested (example data are given line by line). From the measured data, optimal weights are calculated to best predict (rank error minimization) antibiotic efficacy (last column) for the rest of the library (or a new series of similar compounds). This is based on the other criteria, taking into account toxicity as an important negative criterion (weights given in the bottom line).

ing the best ranking are calculated for each parameter based on a series of compounds. Further potential drug targets or putative drugs for a given target can then be ranked and their efficacy predicted considering the same parameters using the optimized weights. For example, one can use established antibiotic compounds from a substance library as standards and the pharmacological parameters of choice — such as half-life, hydrophobicity, N-terminal charge, toxicity — in order to compare and predict new compounds for their antibiotic efficacy based on these parameters (Table 1). This approach is useful for a simple initial analysis and ranking among compounds. However, several further refinements are possible, such as sequential projection pursuit to detect interesting clusters in the multidimensional data available for the compounds measured (Guo et al., 2000).

4. Conclusions

Target identification from genome sequence is achieved here by a combination of sequence comparison, genome context, basic data mining and motif recognition. Such combined approaches also do well on other genome studies and reduce the fraction of unannotated proteins by about a third (Dandekar et al., 2000; Iliopoulos et al., 2000; Koonin, 2001). Evaluation of concrete targets includes genetic algorithm-evaluation of ranking criteria. Potential pharmacological targets are analyzed structurally by domain assignment, homology modeling and model building (RMSD between 6 Å and more, down to 1 Å depending on target difficulty). Three typical examples demonstrate that biologically relevant targets are readily identified. Current procedures are only semiautomatic, relying also on human expert knowledge and intervention. For a number of tasks (e.g. data mining, target ranking), artificial intelligence methods have already been shown to be powerful to retrieve and analyze individual proteins (Casadio et al., 2000; King et al., 2001). Several steps were outlined where we explore and incorporate input from such approaches. Nevertheless, more improvements are expected and necessary (Tsoka and Ouzounis, 2000) to further automate genome target identification in the future.

Acknowledgements

Support by Deutsche Forschungsgemeinschaft (DFG SFB 544/B2; DFG-SPP Bo 1033/5) and the Bundesministerium fuer Bildung und Forschung (project Genominfo) is gratefully acknowledged.

References

Apweiler, R., et al., 2001. Nucleic Acids Res. 29, 37.
Aravind, L., Koonin, E.V., 1999. J. Mol. Biol. 287, 1023.
Bates, P.A., Jackson, R.M., Sternberg, M.J., 1997. Proteins Suppl. 1, 59.

Bernstein, H.J., 2000. Trends Biochem. Sci. 25, 453.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., Yuan, Y., 1998. J. Mol. Biol. 283, 707.

Casadio, R., Compiani, M., Fariselli, P., Jacoboni, I., Martelli, P.L., 2000. SAR QSAR Environ. Res. 11, 149.

Cole, S.T., et al., 1998. Nature 393, 537.

Dandekar, T., et al., 2000. Nucleic Acids Res. 28, 3278.

Emilien, G., Ponchon, M., Caldas, C., Isacson, O., Maloteaux, J.M., 2000. QJM 93, 391.

Fitch, W.M., 1970. Syst. Zool. 19, 99.

Frishman, D., Argos, P., 1997. Proteins 27, 329.

Ghuloum, A.M., Sage, C.R., Jain, A.N., 1999. J. Med. Chem. 42, 1739.

Guo, Q., Questier, F., Massart, D.L., Boucon, C., de Jong, S., 2000. Anal. Chem. 72, 2846.

Helma, C., Gottmann, E., Kramer, S., 2000. Stat. Methods Med. Res. 9, 329.

Huynen, M.A., Dandekar, T., Bork, P., 1998. FEBS Lett. 426, 1.

Iliopoulos, I. et al., 2000. Genome Biol 2, INT1.

King, R.D., Srinivasan, A., Dehaspe, L., 2001. J. Comput.-Aided Mol. Des. 15, 173.

King, R.D., Karwath, A., Clare, A., Dehaspe, L., 2000. Yeast 17, 283.

Koonin, E.V., 2001. Curr. Biol. 11, R155.

Kyrpides, N.C., Ouzounis, C.A., Iliopoulos, I., Vonstein, V., Overbeek, R., 2000. Nucleic Acids Res. 28, 4573.

Levin, J.M., 1997. Protein Eng. 10, 771.

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., Chothia, C., 2000. Nucleic Acids Res. 28, 257.

MacCallum, R.M., Kelley, L.A., Sternberg, M.J., 2000. Bioinformatics 16, 125.

Ota, M., Kawabata, T., Kinjo, A.R., Nishikawa, K., 1999. Proteins 37 (S3), 126.

Pearson, W.R., 2000. Methods Mol. Biol. 132, 185.

Pena-Reyes, C.A., Sipper, M., 2000. Artif. Intell. Med. 19, 1.

Persson, B., 2000. EXS 88, 215.

Pillardy, J., et al., 2001. Proc. Natl. Acad. Sci. USA 98, 2329.

Polanski, J., Walczak, B., 2000. Comput. Chem. 24, 615.

Ramakrishna, R., Srinivasan, R., 1999. Comput. Chem. 23, 165.

Rost, B., Sander, C., 2000. Methods Mol. Biol. 143, 71.

Saxena, I., Brown, R.M., Dandekar, T., 2001. Phytochemistry, in press.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., Bork, P., 2000. Nucleic Acids Res. 28, 231.

Schwede, T., Diemand, A., Guex, N., Peitsch, M.C., 2000. Res. Microbiol. 151, 107.

Tsoka, S., Ouzounis, C.A., 2000. FEBS Lett. 480, 42.

Trohalaki, S., Gifford, E., Pachter, R., 2000. Comput. Chem. 24, 421.