Powers and Pitfalls in Sequence Analysis: The 70% Hurdle

Peer Bork[1]

European Molecular Biology Laboratory (EMBL) 69012 Heidelberg; Germany and Max-Delbrück-Centrum, D-13122 Berlin-Buch, Germany

High-throughput technologies impress us almost every week with novel global results and big numbers. They often reveal important general trends that are impossible to realize with classical, low-throughput experimental methods, yet (so far) they provide fewer insights into specific, molecular detail. Because of the amount of data involved, high-throughput technologies imply the use of bioinformatics methods that deal with information transformation, storage, and analysis. By necessity, most of these processes are automated.

Partly because of the nature of current publication schemes, the accuracy and error margins of a given method are often only found in small print. It is obvious that each method has its limits and also that during data processing, some information will be lost or diluted. Because of the current need to integrate and add value to data, results from high-throughput experiments (if made publicly accessible) are often taken further by third-party research that relies on the quality of these data. Thus, I believe that public awareness of error margins for high-throughput experimental and computational methods should be increased; the incredibly valuable data accumulating in various heterogeneous databases permit powerful analyses but should not be overinterpreted. In the following discussion, I will concentrate on limits in computational sequence analysis, which is far from being perfect (Table 1), despite the fact that sequencing itself is highly automated and accurate, and despite the fact that sequence information is described in simple linear terms (using a four-letter alphabet). On average, a 70% accuracy just to predict functional and structural features has to be considered a success (Table 1).

Limitations in the Total Knowledge Base of Protein Function

As these analysis methods are knowledge based, one of the reasons for the inaccuracy is that the quality of data in public sequence databases is still insufficient (e.g., Bork and Bairoch 1996; Bhatia et al. 1997; Pennisi 1999). This is particularly true for data on protein function. Protein function is loosely defined; cellular function is more than the very complicated network of individual molecular interactions on which it is based (Bork et al. 1998). Furthermore, the semantics for functional features are not always established. For instance, the notion of a "protein complex" not only depends heavily on detection and purification methods—which, in turn, are constantly evolving—but also on environmental conditions. Protein function is context dependent, and both molecular and cellular aspects have to be considered (for review, see Bork et al. 1998).

To illustrate some of this complexity, a good example is lactate dehydrogenase: This gene product can act both as a dehydrogenase and an eye lens structural protein, depending on its context (for review, see Piatigorsky and Wistow 1991). Even without the complication of a second, unrelated role for the same gene product, do we know enough about the function of lactate dehydrogenase, one of the best-studied proteins? We know its biochemical pathway (at least in human and some model organisms), its different isoenzymes (in organisms) with different context-dependent properties, its regulation, and the organization of its quaternary structure. However, we are probably still missing much information, even on crucial molecular features: Are we sure about alternative splice variants? Can we exclude age-dependent post-translational modifications in some tissues? Our knowledge is even more limited regarding higher order functions that involve concentration, compartmental organization, dynamics, regulation, and perhaps even the impact of external environment. Often, the available data give at best some reliable qualitative results on functional features but far from a complete understanding of functionality. Yet our ability to annotate genome sequences and translate information therein relies heavily on the summaries of features attached to each sequence in the respective public databases.

Limitations of Gene Expression Data Extrapolations

As more high-throughput technologies follow, the data will become more complicated than sequences. Novel complementary data types such as gene expression arrays will generate more functional information, but conclusions from these data are often stretched with regard to protein products. The expression of genes and their reciprocal proteins seems to correlate weakly, with a correlation coefficient of 0.48 (Anderson and Seilhammer 1997). Furthermore, recent studies (Hanke et al. 1999; Mironov et al. 1999) show that alternative splicing might affect >30% of the human genes, although measurements at the protein level have yet to confirm this. Finally, the number of known post-

[1] **E-MAIL bork@embl-heidelberg.da; FAX 11-49-6221-387517.**

10:398–400 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

Table 1. Selected Examples of Prediction Accuracy in Different Areas of Sequence Analysis

Prediction of	Acc × cov[a]	Accuracy (%)	Coverage or coverage in % of reference set	Reference[b]
Human promoters	0.35	50	70% of annotated test set	Prestidge 1995; P. Bucher (pers. comm)
Human regulatory RNA elements	0.34	85	40% of new DNA	Dandekar and Sharma (1998)
Human genes (only presence)	0.49	70	70% of chromosome 22	Dunham et al. (1999) and refs. therein
Human SNPs by EST comparison	0.21	70	30% of all proteins with SNP	Buelow et al. (1999); Sunyaev et al. (2000)
Human alternative splicing	0.45	90	50% of all splice sites	Hanke et al. (1999)
Transmembranes (only presence)	0.85	85	99% of annotated test set	Tusnady and Simon (1998) and refs. therein
Signal peptides (only presence)	.90	90	100% of annotated test set	Nielsen et al. (1999)
GPI ancors (incl cleavage site)	.72	72	100% of annotated test set	Eisenhaber et al. (1999)
Coiled coil (only presence)	.81	90	90% of annotated coiled coil	Lupas (1996)
Secondary structure (Three states)	.77	77	100% of 3D test set	Jones (1999) and refs. therein
Buried or exposed residues	.74	74	100% of 3D test set	Rost (1996)
Residue hydration	.72	72	100% of 3D test set	Ehrlich et al. (1998)
Protein folds (in Mycoplasma)	.49	98	50% of Mycoplasma ORFs	Teichmann et al. (1999) and refs. therein
Homology (several methods)	.49	98	50% of 3D test set	Muller et al. (1999) and refs. therein
Functional features by homology	.63	90	70% unicellular genomes	Bork and Koonin (1998); Brenner (1999)
Function association by context	.25	50	10% high confidence in yeast	Marcotte et al. (1999b)
Cellular localization (two states)	.77	77	100% of annotated test set	Andrade et al. (1998)

The numbers referred to are in many cases crude estimates taken or sometimes even estimated from the literature and have an expected accuracy of ~70%. Direct comparison of the numbers might be misleading as the context is not properly explained here. Furthermore, although most of the examples are two state predictions, the percentage numbers do not take into account random occurrences of the states. All test sets are most likely biased (e.g., current 31) test sets do not contain many compositionally biased regions, which probably contain up 15% of all residues, and annotation test sets are far from being perfect; see text), i.e., the real accuracy is thus probably lower.

[a] To make the numbers more comparable, accuracy has been multiplied by coverage; some methods give accuracy for different degree of coverage and roughly justify this procedure. However, often it is biased toward sensitivity as specificity cannot be properly taken into account. Most features predicted with an accuracy × coverage >0.70 are of structural nature and at best only indirectly imply a certain functionality.

[b] Only one recent reference is given and if indicated, references therein should also be considered as other reports do not always agree with the numbers given.

translational modifications of gene products is increasing constantly, so that the complexity at the protein level is enormous. Each of these modifications may change the function of the respective gene products drastically. (The entire aspect of context-dependent gene regulation is excluded from current discussions as we are only beginning to understand the complex underlying genetic machinery. For example, promoter prediction in eukaryotes has a success of only ~35% (Table 1), and there are many other regulatory elements that we cannot predict at all.)

Limitations Created by Third-Party Analyses

Public releases of completely sequenced genomes exceed a rate of one per month, with thousands of function predictions therein. Gene annotation via sequence database searches is already a routine job, but even here the error rate is considerable (Table 1). The lower limit of errors in current functional annotation of large-scale sequencing projects is 8% (Brenner 1999). As errors accumulate and propagate (Bork and Bairoch 1996; Bhatia et al 1997; Smith and Zhang 1997; Bork and Koonin 1998; Pennisi 1999), it becomes more difficult to infer correct function from the many possibilities revealed by a database search. Increasing these complications is the fact that computer programs often cannot even retrieve the source of the stored information (Doerks et al. 1998).

Use of Complementary Information to Limit Errors in Function Prediction

Some new information can be retrieved from completely sequenced genomes, for example, function can be predicted by exploitation of genomic context.

Based on the observation that interacting proteins in one organism sometimes have homologs in other organisms fused together in a single gene, Marcotte et al. (1999a) predicted novel interactions for 50% of yeast proteins using gene fusion information. However, they noted an overlap with classical methods and an error rate of 82%. To see a signal they had to correct for domains present in many proteins (Marcotte et al. 1999a). By considering only orthologs with fission and fusion events (Enright et al. 1999, Snel et al. 2000), the signal-to-noise ratio increases and the number of predictions drops dramatically (7% of *Escherichia coli* proteins; Enright et al. 1999). With a particular question in mind, Does protein X have interaction partners?, the generation of hypotheses is extremely useful; yet to provide a general overview of protein function, it is advisable to keep the errors small. Further information can be added later, which is easier than retracting stored information. But how do we incorporate the information on error margins? Such estimates (sometimes not even the sources of the annotation) are not visible in current databases that store the results of computational approaches.

Taking the 70% Hurdle

As noted above, most prediction schemes extrapolate from current knowledge, and many bioinformatics methods have difficulty exceeding a 70% prediction accuracy (numbers in Table 1 are often overestimates because the test sets used are usually not representative of all sequences). On one hand, current methods seem to capture important features and explain general trends; on the other hand, 30% of the features are missing or predicted wrongly. This has to be kept in mind when processing the results further. Also the 70% accuracy often attaches to methods that deal with discrete objects such as sequences; making estimates about the prediction of cellular features is much more difficult as one first has to agree on semantics (or ontology in a database sense) to describe complex processes in a comparable way.

All of the above focuses on limitations in the computational prediction of qualitative features. There remains a long way to go until we are able to describe molecular processes quantitatively; current simulations of complex systems are still very rough and simplistic. However, there is still no doubt that sequence analysis is extremely powerful and that the generation of hypotheses derived by computational methods will be more and more often the first successful step in the design of experiments. If 70% of such experiments were successful, the speed of scientific discoveries would grow exponentially.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Anderson, L. and J. Seilhammer. *Electrophoresis* **18:** 533–537.

Andrade, M., S.I. O’Donoghue, and B. Rost. 1998. *J. Mol. Biol.* **276:** 517–525.

Bhatia, U., K. Robison, and W. Gilbert. 1997. *Science* **276:** 1724–1725.

Bork, P. and A. Bairoch. 1996. *Trends Genet.* **12:** 425–427.

Bork, P. and E.V. Koonin. 1998. *Nat. Genet.* **13:** 313–318.

Bork, P., T. Dondekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan. 1998. *J. Mol. Biol.* **283:** 707–725.

Brenner, S. 1999. *Trends Genet.* **15:** 132–133.

Buelow, K.H., M.N. Edmonson, and A.B. Cassidy. 1999. *Nat. Genet.* **21:** 323–325.

Dandekar, T. and K. Sharma. 1998. *Regulatory RNA*. Springer Verlag, Heidelberg, Germany.

Doerks, T., A. Bairoch, and P. Bork. 1998. *Trends Genet.* **14:** 248–250.

Dunham, I., N.Shimizu, B.A. Roe, S. Chissoe, J.E. Colllins, R. Bruskiewich, M. Clamp, L.J. Smink, R. Ainscough, and J.P. Almeida. 1999. *Nature* **402:** 489–495.

Ehrlich L, M. Reczko, H. Bohr, and R.C. Wade. 1998. *Protein Eng.* **11:** 11–19.

Eisenhaber, B., P. Bork, and F. Eisenhaber. 1999. *J. Mol. Biol.* **292:** 741–758.

Enright, A.J. I.Iliopoulos, N.C. Kyrpides and C.A. Ouzounis. 1999. *Nature* **402:** 86–90.

Hanke, J., I. Zastrow, A. Aydin, G. Lehmann, S. Luft, J.G. Reich, and P. Bork. 1999. *Trends Genet.* **15:** 389–390.

Jones, D.T. 1999. *J. Mol. Biol.* **292:** 195–202.

Lupas, A. 1996. *Methods Enzymol.* **266:** 513–525.

Marcotte E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. 1999a. *Science* **285:** 751–753.

Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999b. *Nature* **402:** 83–86.

Mironov, A.A., J.W. Fickett, and M.S. Gelfand. 1999. *Genome Res.* **15:** 755–771.

Muller, A., R.M. MacCallum, and M.J.E. Sternberg. 1999. *J. Mol. Biol.* **293:** 1257–1271.

Nielsen, H., S. Brunak, and G. von Heijne. 1999. *Protein Eng.* **12:** 3–9.

Pennisi, E. 1999. *Science* **286:** 447–450.

Piatigorski, Y. and G.J. Wistow. 1991. *Science* **252:** 1078–1079.

Prestidge, D.S. 1995. *J. Mol. Biol.* **249:** 923–932.

Rost, B. 1996. *Methods Enzymol.* **266:** 525–539.

Smith, T.F. and X. Zhang. 1997. *Nat. Biotechnol.* **15:** 1222–1223.

Snel, B., P. Bork, and M. Huynen. 2000. *Trends Genet.* **16:** 9–11.

Sunyaev, S., J. Hanke, D. Brett, A. Aydin, I. Zastrow, W. Lathe, P. Bork and J. Reich. 2000. *Adv. Protein Chem.* **54:** (in press).

Teichmann, S., C. Chothia, and M. Gerstein. 1999. *Curr. Opin. Struct. Biol.* **9:** 390–399.

Tusnady, G.E. and I. Simon. 1998. *J. Mol. Biol.* **283:** 489–506.