BIOINFORMATICS DISCOVERY NOTE

Vol. 15 no. 9 1999
Pages 759–762

Finding families for genomic ORFans

Daniel Fischer[1,*] and David Eisenberg[2]

[1]Faculty of Natural Science, Department of Math and Computer Science, Ben Gurion University, Beer-Sheva 84015, Israel and [2]UCLA-DOE Laboratory of Structural Biology & Molecular Medicine, Molecular Biology Institute, UCLA, BOX 951570, Los Angeles, CA 90095-1570, USA

Received March 10, 1999; revised May 10, 1999; accepted May 12, 1999

'Why, if species have descended from other species by insensibly fine gradations, do we not everywhere see innumerable transitional forms?'

Charles Darwin, in *The Origin of Species*, Chapter 6: **Difficulties in Theory**

The complete sequences of over a dozen microbial genomes are now known. At first glance roughly one-third of the protein encoding regions (ORFs) in each genome have no detectable sequence similarity to proteins of other genomes. Why, if proteins in different organisms have descended from common ancestral proteins by duplication and adaptive variation, do so many today show no similarity to each other? In this commentary we refer to these orphan ORFs as 'ORFans' and ask why there are so many, and how they can be assigned to known protein families. As a first step in solving the puzzle of why there are so many genomic ORFans, we have investigated some trivial explanations.

Trivial reasons for many ORFans

The first trivial explanation is that the number of ORFans reported by the original authors (given in the 'Original ORFans' column of Table 1) are outdated because these initial ORFans may now find matches with genes sequenced more recently. To check this, we recomputed the number of ORFans in each genome using a standard sequence comparison method (with standard parameters; Altschul *et al*., 1990)†, comparing each ORF originally annotated as an ORFan against the sequences from all the completed genomes and all the protein sequences in SWISSPROT. The results in the column labeled 'Updated ORFans' show that the fraction of ORFans has diminished, but is still significantly high. Furthermore, the percentage of ORFans observed in the more recently determined genomes demonstrates that the fraction of ORFans is not diminishing. Only for the genomes of closely related organisms is the fraction of ORFans expected to drop significantly. For example, all of the ORFans of *M. genitalium* now have homologs in *M. pneumonia*, but the question remains of why numerous proteins in one group of closely related organisms do not match proteins in other groups of organisms.

A second trivial possibility for finding many ORFans involves near duplication of ORFans in their own genomes. That is, two or more ORFans in the same genome may belong to the same paralog family. These near duplicates all count as ORFans because they have no matches in other organisms. However, for our quest of understanding why there are so many ORFans, it is more meaningful to compute the number of ORFan paralogous families in each genome (next column in the table) than the total number of ORFans. The reduction from total ORFans to ORFan paralogous families varies depending on the extent of duplication that each genome has experienced but still leaves many ORFan paralogous families. As an aside, we note that ORFans have fewer paralogs than other ORFs. That is, sequences without relatives in other genomes are less likely to have near duplicates in their own genome. Does this suggest that many ORFans are new proteins that have not had enough time to duplicate within the organism? And if so, where did they come from? Alternatively, could it be that ORFans are 'dead-end' sequences, i.e. after duplication and modification, ORFans never produce proteins with any functional advantage, and thus disappear?

A third possible trivial explanation is that many ORFans may correspond to 'junk' DNA, that is, segments of DNA identified as ORFs but not expressed as proteins (Goffeau *et al*., 1996; Dujon *et al*., 1994). These junk ORFans may have mutated from non-coding segments of DNA, may be descendants of genes that are no longer functional (e.g. pseudogenes, dead genes or remnants of horizontal transfer) or may correspond to wrongly predicted ORFs. These junk ORFans might continue to mutate until they cease to appear as ORFs or in rare cases until they become new functional proteins. In either case, it is unlikely that junk ORFans would match other ORFs, and should be excluded from our count. To minimize junk ORFans we focused on the larger ORFs only (column 'All' in the table), based on the observation that junk ORFans are likely to be short (Dujon *et al*., 1994). The percentages of ORFans among longer ORFs are lower than those shown in the previous column, that is, ORFans are more frequent among shorter ORFs than in the genome as a whole, but still a significant number of ORFans remain.

*Corresponding author. E-mail: dfischer@cs.bgu.ac.il

†We use a standard sequence comparison method not only to be consistent with previous genome annotations but also because its sensitivity exactly fits our notion of ORFan: an ORF with no **clearly detectable** sequence similarity to other proteins.

Table 1. ORFans from 8 microbial genomes

Organism	Ref	No. ORFs	Original[a] ORFans	Updated[b] ORFans	ORFan paralogous families[c]			
					Total	>150 residues		
						All	Soluble	
H. influenzae	1	1743	22%	12%	11%	6%	6%	59
M. genitalium	2	470	20%	0%	0%	0%	0%	0
M. jannaschii	3	1738	56%	25%	21%	16%	14%	146
M. pneumonia	4	677	16%	11%	7%	6%	4%	18
Synechoscystis sp.	5	3168	45%	33%	29%	24%	22%	406
H. pylori	6	1590	31%	29%	25%	18%	18%	172
A. fulgidus	7	1855	27%	16%	14%	10%	8%	110
B. burgdorferi	8	852	29%	26%	26%	22%	20%	99

References: (1) *Science*, **269,** 496–512, 1995; (2) *Science*, **270,** 397–403, 1995; (3) *Science*, **273,** 1058–1073, 1996; (4) *Nucleic Acids Res*., **24,** 4420–4449, 1996; (5) *DNA Res*., **3,** 109–136, 1996; (6) *Nature*, **388,** 539–547, 1997; (7) *Nature*, **390,** 364, 1997; (8) *Nature*, 390:580–586, 1997.

[a]Computed by original authors from other sequences available at the time of publication.
[b]Computed by the present authors from all genomes available as of December 20, 1997 and all SWISSPROT sequences (see also other updated annotations at http://www.ncbi.nlm.nih.gov/Complete_Genomes or at http://www.sander.ebi.ac.uk/genequiz).
[c]An ORFan paralogous family is a family of one or more sequences in a genome whose members match no ORFs in another genome. **Total** indicates the number of ORFan paralogous families divided by the number of ORFs in the genome (column 3) times 100. **All** indicates the number of ORFan paralogous families whose ORFs are longer than 150 residues divided by the total number of ORFs in the genome longer than 150 residues times 100. **Soluble** indicates the number of soluble ORFan paralogous families longer than 150 residues divided by the total number of soluble ORFs in the genome longer than 150 residues times 100. The last number is the total number of soluble, ORFan paralogous families longer than 150 residues. Non-soluble proteins were identified with the program MOMENT (*J. Mol. Biol*., **179,** 125–142, 1984).
The sequenced genomes not included in the table (but used in the update) are: *E. coli* (*Science*, **277,** 1453–1462, 1997), *B. Subtilis* (*Nature*, **390,** 249–256, 1997), Yeast (*Nature* **387,** suppl. 5-105, 1997), *M. thermoautotrophicum* (*J. Bacteriology*, 7135–7155, 1997) and a preliminary version of *Pyrobaculum aerophilum* (*Extremophiles*, **1,** in press, 1997).

A fourth trivial reason for many ORFans is that they may represent membrane proteins. Our knowledge of membrane proteins lags behind that of soluble proteins in many ways, including how to recognize distant family members. Because of this, we removed all ORFs containing putative trans-membrane helical regions and found that putative membrane proteins appear slightly more frequently among the ORFans than in the genomes as a whole. But even after removing membrane proteins from the ORFans, there remains a substantial fraction of ORFans in most genomes, as shown in the final two columns of the table.‡

Are ORFans unique proteins or undetected members of known protein superfamilies?

The numbers of 'ORFan paralogous families' listed in the final column of the table are the number of longer, soluble protein families that do not have detectable sequence similarity to any protein in other genomes. There are at least two less trivial explanations for these still sizable numbers of ORFans.

The first is that they are not true orphans but rather are distant members of known protein superfamilies that standard sequence comparison methods fail to detect. That is, through adaptive variation ORFans have diverged beyond recognition at the sequence level, but their functions and 3D structures are similar to known superfamilies. In this case, the estimates of one (Chothia, 1992) or a few (Orengo *et al*., 1994) thousand superfamilies in nature seem to be of the correct order of magnitude, and the ever growing number of ORFans merely reflects our present inability to assign them to their proper families. To detect such distant relationships more sensitive bioinformatics methods are needed. Higher sensitivity can be obtained using evolutionary information inherent in multiple sequences (e.g. Gribskov *et al*., 1987; Krogh *et al*., 1994; Altschul *et al*., 1997; Karplus *et al*., 1998) or by incorporating three-dimensional information (e.g. fold assignment methods; Bowie *et al*., 1991; Jones *et al*., 1992; Casari and Sippl, 1992; Fischer and Eisenberg, 1997).

More sensitive sequence methods (e.g. Altschul *et al*., 1997; Karplus *et al*., 1998) succeed in finding distant relationships for many of the non-ORFans. But for ORFans, they succeed

‡We also examined whether the presence of non-globular proteins might account for the large number of ORFans, but only four of the ORFans in our final count appeared to lack a globular domain (they showed no contiguous segment of size 60 containing no low-complexity regions).

only in a small number of cases (results not shown). The increased sensitivity of these methods stems mainly from their incorporation of evolutionary information from neighboring sequences and because by definition, ORFans have no sequence neighbors, these methods find distant relationships for only a few of the ORFans.

Fold-assignment methods can be helpful in assigning families (via 3D folds) for genomic ORFans. The three-dimensional information helps detect distant relationships because 3D folds evolve more slowly than sequences. For example ORF MG353 from *M. genitalium* was originally classified as an ORFan because searches against the database available at that time revealed no significant similarity to any sequence. However, fold assignment (Fischer and Eisenberg, 1997) identified the structure of a histone-like fold (PDB code 1hue) as highly compatible with MG353, permitting the former ORFan MG353 to be assigned to this superfamily of DNA-binding proteins. In short, fold assignment methods can establish some distant protein relationships that standard sequence comparison methods cannot. However, these methods are not yet powerful enough to assign many ORFans to known protein families (Fischer and Eisenberg, 1997) and work only when the 3D structure is known for at least one family member.

Even if the vast majority of ORFans correspond to undetected members of known protein superfamilies, the question of how much diversity exists at the sequence level remains. That is, why is it that the sequences of these ORFans are so distant from their relatives in other organisms? Why have these ORFans diverged so far that no close relatives are observed today?

A second explanation for the large number of ORFans is that the ORFans code for proteins unique to an organism, or to a closely related group of organisms. They do not match proteins in other species because they belong to different protein superfamilies with unique functions and structures. In other words, these ORFans may be the determinants of the species. But where have they come from? Through adaptive descent the sequences of ORFans may have diverged from ancestral proteins to such an extent that they have acquired new, unique functions and structures not observed in any other organisms.

Alternatively, ORFans may reflect gene-loss in other organisms. This proposed origin of ORFans in gene-loss implies that these proteins were once present in many organisms, but during evolution, species have disappeared leaving ORFan genes in only one or a small group of closely related organisms. That is, present day ORFans may be the only descendants of a once common line of proteins. They still survive in a given species because they fulfill a specific function or because they simply have not had enough time to disappear.

In this view, the function and structure of an ORFan are unique and thus each ORFan defines a newly discovered superfamily. If ORFans continue to appear in each newly sequenced genome, then the number of protein superfamilies may be larger than earlier estimates (Chothia, 1992; Orengo *et al*., 1994). For ORFans corresponding to new superfamilies, bioinformatics cannot find distant relationships to known superfamilies because there are none. Nevertheless, these ORFs are likely to be among the most interesting targets for further structural and functional studies.

The growing number of ORFans and the number of sequence families in nature

Whether ORFans are distant relatives of known superfamilies with similar functions and structures, or are proteins with new functions and structures, unseen to date in other organisms — the puzzle remains of why there are many of them and why they lack close sequence relatives. We note that Darwin commented on a similar phenomenon at the species level in the Origin of the Species: '*... numberless intermediate varieties... must assuredly have existed; but the very process of natural selection, constantly tends... to exterminate the parent-forms and the intermediate links*'. Unless the fraction of ORFans significantly drops in the next round of sequenced genomes, the number of sequence families detectable by sequence similarity ('30SEQ' families of Orengo, 1994), will be larger than previously estimated (23 100), and the sequence databases will increasingly fill with new sequences awaiting characterization.

The role of bioinformatics and structural biology in interpreting genomes

Structural biology will be essential in determining whether ORFans are the first examples of unique protein superfamilies, or are distant members of known protein superfamilies. Bioinformatics can play an important role in identifying potentially interesting ORFans, aiding the priorization of their characterization, as in structural genomic initiatives. If an ORFan represents a unique superfamily it is important to learn its 3D structure as part of characterizing its function. If an ORFan instead is a distant member of a known superfamily, its structure may be required to determine its relationship. In any case, until the 3D structures of ORFans are experimentally determined, more sensitive bioinformatics tools will aid in placing genomic ORFans into their proper protein superfamiles.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment tool. *J. Mol. Biol*., **215,** 403–410.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res*., **25,** 3389–3402.

Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253,** 164–170.

Casari,G. and Sippl,M.J. (1992) Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.*, **224,** 725–732.

Chothia,C. (1992) One thousand families for the molecular biologist. *Nature*, **357,** 543–544.

Dujon,B. *et al*. (1994) Complete DNA sequence of yeast chromosome XI. *Nature*, **369,** 371–377.

Fischer,D. and Eisenberg,D. (1997) Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl Acad. Sci. USA*, **94,** 11929–11934.

Goffeau,A., Barrell,B.G., Bussey,H., Davids,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., Louis,E.J., Mewes,H.W., Murakami,Y., Philippsen,P., Tettelin,H. and Oliver,S.G. (1996) Life with 6000 genes. *Science*, **274,** 546–547.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: Detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84,** 4355–4358.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature*, **358,** 86–89.

Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden markov models for detecting remote protein homologies. *Bioinformatics*, **14,** 846–856.

Krogh,A., Brown,M., Mian,I.S., Solander,K. and Haussler,D. (1994) Hidden markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, **235,** 1501–1531.

Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) Protein superfamilies and domain superfolds. *Nature*, **372,** 631–634.