Structural Biology Sheds Light on the Puzzle of Genomic ORFans

Naomi Siew¹,²,³ and Daniel Fischer²,³*

Genomic ORFans⁴ are orphan open reading frames (ORFs) with no significant sequence similarity to other ORFs. ORFans comprise 20–30% of the ORFs of most completely sequenced genomes. Because nothing can be learnt about ORFans via sequence homology, the functions and evolutionary origins of ORFans remain a mystery. Furthermore, because relatively few ORFans have been experimentally characterized, it has been suggested that most ORFans are not likely to correspond to functional, expressed proteins, but rather to spurious ORFs, pseudo-genes or to rapidly evolving proteins with non-essential roles. As a snapshot view of current ORFan structural studies, we searched for ORFans among proteins whose three-dimensional structures have been recently determined. We find that functional and structural studies of ORFans are not as underemphasized as previously suggested. These recently determined structures correspond to ORFans from all Kingdoms of life, and include proteins that have previously been functionally characterized, as well as structural genomics targets of unknown function labeled as “hypothetical proteins”. This suggests that many of the ORFans in the databases are likely to correspond to expressed, functional (and even essential) proteins. Furthermore, the recently determined structures include examples of the various types of ORFans, suggesting that the functions and evolutionary origins of ORFans are diverse. Although this survey sheds some light on the ORFan mystery, further experimental studies are required to gain a better understanding of the role and origins of the tens of thousands of ORFans awaiting characterization.

Keywords: genomic ORFans; evolution; structural biology

¹Department of Chemistry
Ben Gurion University
Beer-Sheva 84105, Israel
²Bioinformatics Group,
Department of Computer Science, Ben Gurion University
Beer-Sheva 84105, Israel
³Buffalo Centre of Excellence in Bioinformatics/Computer Science and Engineering
University at Buffalo, 901 Washington St, Suite 300
Buffalo, NY 14203, USA

*Corresponding author
E-mail address of the corresponding author:
dfischer@bioinformatics.buffalo.edu

Abbreviations used: ORF, open reading frame; ORFans, orphan ORFs; FR, fold-recognition.

Present address: Both authors, Buffalo Center of Excellence in Bioinformatics/Computer Science and Engineering, University at Buffalo, 901 Washington St. Suite 300, Buffalo, NY 14203, USA.
novel proteins, unique to an organism or a lineage, with possibly new functions and/or 3D structures.\(^1,3\) In either case, the mystery of why ORFans have no homologs remains. Are ORFans the result of rapid evolution,\(^4,15,16\) of lateral gene transfer\(^17\) from unknown organisms or are they the result of gene-losses or of de novo generation?\(^5,6,18-20\) On the other hand, it has also been suggested that most ORFans, especially the shorter ones, may correspond to non-essential, non-functional or non-expressed proteins.\(^15,21-24\)

Here, as a snapshot view of current ORFan structural studies, we report a recent survey we conducted among newly determined 3D protein structures, and show how structural biology is already being essential in unraveling the ORFan puzzle.\(^1,10,25\) We searched for ORFans among the PDB\(^26\) entries released between June and December 2003 and found that out of the 172 protein chains sharing no significant sequence similarity to previously determined protein structures,\(^27,17\) 17 correspond to ORFans and two correspond to "poorly conserved ORFs" or POCOs\(^2\) (Table 1; for simplicity, in what follows we refer to these 19 proteins as ORFans; see below). This strongly suggests that many ORFans correspond to real, foldable proteins, and not to sequencing errors or dead proteins. The relatively large percentage of ORFans among the newly determined structures (11\%) suggests that ORFans may not be as underemphasized as previously suggested\(^1,2\) and that experimental studies of ORFans have already become routine.

In what follows, we show that these 19 ORFans provide interesting examples of the various types of ORFans.

Thirteen of these 19 newly determined ORFans correspond to proteins whose function was previously characterized experimentally (at least in the broad sense), and thus, are not "orphans" with regards to their functions. This suggests that many more ORFans with still unknown 3D-structures have already been characterized functionally. These 13 ORFans cover various functional categories, with at least five involved in transcription/translation, suggesting that, because of the high sequence divergence required in these processes, many ORFans may belong to these categories. The other six ORFans correspond to proteins of unknown evolution,\(^4,15,16\) of lateral gene transfer 17 from unknown organisms or are they the result of gene-losses or of de novo generation?\(^5,6,18-20\) On the other hand, it has also been suggested that most ORFans, especially the shorter ones, may correspond to non-essential, non-functional or non-expressed proteins.\(^15,21-24\)

Here, as a snapshot view of current ORFan structural studies, we report a recent survey we conducted among newly determined 3D protein structures, and show how structural biology is already being essential in unraveling the ORFan puzzle.\(^1,10,25\) We searched for ORFans among the PDB\(^26\) entries released between June and December 2003 and found that out of the 172 protein chains sharing no significant sequence similarity to previously determined protein structures,\(^27,17\) 17 correspond to ORFans and two correspond to "poorly conserved ORFs" or POCOs\(^2\) (Table 1; for simplicity, in what follows we refer to these 19 proteins as ORFans; see below). This strongly suggests that many ORFans correspond to real, foldable proteins, and not to sequencing errors or dead proteins. The relatively large percentage of ORFans among the newly determined structures (11\%) suggests that ORFans may not be as underemphasized as previously suggested\(^1,2\) and that experimental studies of ORFans have already become routine.

In what follows, we show that these 19 ORFans provide interesting examples of the various types of ORFans.

Thirteen of these 19 newly determined ORFans correspond to proteins whose function was previously characterized experimentally (at least in the broad sense), and thus, are not "orphans" with regards to their functions. This suggests that many more ORFans with still unknown 3D-structures have already been characterized functionally. These 13 ORFans cover various functional categories, with at least five involved in transcription/translation, suggesting that, because of the high sequence divergence required in these processes, many ORFans may belong to these categories. The other six ORFans correspond to proteins of unknown

**Table 1. The 19 ORFans with a recently determined 3D structure**

<table>
<thead>
<tr>
<th>PDB code</th>
<th>Organism</th>
<th>PDB description</th>
<th>Length (aa)</th>
<th>Has homologs in</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ofSB</td>
<td><em>S. cerevisiae</em></td>
<td>Mrna export factor Mtx67-Mtx2</td>
<td>184</td>
<td>-</td>
</tr>
<tr>
<td>1q1A</td>
<td><em>S. solitarius</em></td>
<td>Hypothetical protein Apc1120 (protein Yesu)</td>
<td>223</td>
<td>-</td>
</tr>
<tr>
<td>1mw5A</td>
<td><em>H. influenzae</em></td>
<td>Hypothetical protein H1480</td>
<td>187</td>
<td>-</td>
</tr>
<tr>
<td>1ppSU</td>
<td><em>T. vaginalis</em></td>
<td>Initiator binding domain (TBP39)</td>
<td>132</td>
<td>-</td>
</tr>
<tr>
<td>1q7A</td>
<td><em>T. vaginalis</em></td>
<td>C-domain of the Inr binding protein</td>
<td>138</td>
<td>-</td>
</tr>
<tr>
<td>1q77A</td>
<td><em>A. aolicus</em></td>
<td>Putative Universal Stress Protein (Hypothetical Protein Ag_178)</td>
<td>221</td>
<td>-</td>
</tr>
<tr>
<td>1rf8B</td>
<td><em>S. cerevisiae</em></td>
<td>Translation initiation factor Eif4E</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>1nycA</td>
<td><em>S. aureus</em></td>
<td>Staphostatin B</td>
<td>111</td>
<td>S. warneri</td>
</tr>
<tr>
<td>1oh1</td>
<td><em>S. aureus</em></td>
<td>Staphostatin A (hypothetical protein Sav1910)</td>
<td>109</td>
<td>S. epidermidis</td>
</tr>
<tr>
<td>1q8A</td>
<td><em>M. genitalium</em></td>
<td>Conserved hypothetical protein (Mg027)</td>
<td>151</td>
<td>M. pneumoniae</td>
</tr>
<tr>
<td>1osyA</td>
<td><em>F. velutipes</em></td>
<td>Fip-Fve fungal immunomodulatory protein</td>
<td>115</td>
<td>G. lucidum</td>
</tr>
<tr>
<td>1r75A</td>
<td><em>L. major</em></td>
<td>Hypothetical protein</td>
<td>110</td>
<td>T. brucei</td>
</tr>
<tr>
<td>1n1gA</td>
<td><em>T. vulcanum</em></td>
<td>Hypothetical protein (Ta1238)</td>
<td>152</td>
<td>T. acidophilum, F. acidarmanus</td>
</tr>
<tr>
<td>1ofzA</td>
<td><em>A. aurantiaca</em></td>
<td>Fungal lectin</td>
<td>312</td>
<td>A. oryzae, A. fumigatus</td>
</tr>
<tr>
<td>1u2F</td>
<td><em>Rice dwarf virus</em></td>
<td>Rice dwarf virus (capsid protein)</td>
<td>421</td>
<td>Wound tumor virus, Rice gall</td>
</tr>
<tr>
<td>1q6aA</td>
<td><em>T. elongatus</em></td>
<td>Circadian clock protein KaiA homolog</td>
<td>214</td>
<td>T. erythropneum, Nostoc punctiforme, Nostoc sp. PCC 9709, PCC 7120 &amp; P. menenbraca Synechococcus sp. PCC &amp; WH, Synechocystis sp. PCC</td>
</tr>
<tr>
<td>1qv9A</td>
<td><em>M. kaudleri</em></td>
<td>Coenzyme F420-dependent Mtd</td>
<td>283</td>
<td>M. jannaschii, M. mazei, M. acetivorans, M. barkeri, M. thermautotrophicus, M. thermaautotrophicum, A. fulgidus</td>
</tr>
</tbody>
</table>
| 1q5zA    | *S. typhimurium*          | C-terminal actin binding domain of Salmonella invasion protein A (Sipa)          | 177         | S. enteritidis, S. enterica Typhi, S. violaceum
| 1n93A    | *Borna virus*             | Nucleoprotein                                                                    | 375         | H. sapiens, M. musculus

ORFans are ORFs lacking significant sequence similarity with other ORFs, except for possibly ORFs from closely related organisms. We assess significant sequence similarity using the standard PSI-BLAST\(^26\) sequence comparison tool until convergence, and the e-value threshold of 0.001. References for the PDB codes are included only if they were published before March 2004.\(^1\) Because of the presence of a homolog in *Chromobacterium violaceum*, 1q5zA is not a proper orthologous ORFan, but rather a poorly conserved ORF or POCO.\(^2\) 1n93A is also a POCO, which has probably been laterally transferred to humans and mice (see the text).
function (annotated as “hypothetical proteins”) whose structure was determined as part of structural genomics projects. Seven of the 19 ORFans correspond to singleton ORFans, and ten correspond to orthologous ORFans (have homologs only within closely related organisms). The 19 ORFans belong to organisms spanning all kingdoms: seven from Bacteria, three from Archea, seven from Eukarya and two viruses.

Despite the fact that ORFans show no significant sequence similarity to other proteins, the 3D structures of the majority of the ORFans clearly have previously observed folds. This suggests that they either correspond to highly divergent distant members of known protein families, with possibly similar functions or to proteins with unrelated functions whose structures have converged to a similar fold. For example, the essential messenger RNA export factor Mtr2 from yeast (1of5B), a singleton ORFan, was known to be similar in function to the metazoan p15 protein.\(^{28}\) The previously determined 3D structure of the p15 protein revealed that it belongs to the NTF2-like family. The 3D structure of 1of5B revealed that Mtr2 is similar to that of p15, and thus, 1of5B represents a novel member of this family. This is a clear example of an ORFan with a highly divergent sequence, whose function and structure are similar to those of a known family. An example of an ORFan with a previously known fold, but with an unrelated function, is the bacterial virulence factor staphostatin B (1nyCA), a cysteine protease inhibitor with a Staphylococcus-specific function.\(^{29}\) Unexpectedly, its 3D structure turned out not to be similar to other cystatins, but rather, to a variation of the lipocalin fold. This was a surprising result because staphylococci were not expected to contain lipocalin-like functions and no evidence of lipocalin-like properties has been identified in 1nyCA. These examples illustrate that the 3D-structures of those ORFans whose broad function is known, can help to better understand their mechanisms of operation, and in some cases reveal their evolutionary origins.

The 3D-structures of the ORFans of unknown function, if similar to previously observed folds, can also be of help to generate verifiable hypotheses regarding their possible function and/or origin. One such example is the 3D-structure of the singleton ORFan of unknown function from Aquifex aeolicus (1q77A). Its structural similarity to the family of universal stress proteins lead to a putative functional assignment, which, if true, would imply that 1q77A is another highly divergent member of that family.

Could bioinformatics tools have predicted the approximate structures of those ORFans having previously observed folds? Fold-recognition (FR) methods,\(^{30,31}\) applied without using the information from the recently released structures,\(^{27}\) but using as templates previously determined structures, correctly predicted the structures of six of the 19 ORFans. Figure 1 shows the highly confident FR prediction of a Mycoplasma-specific hypothetical protein (1q8cA). This M. genitalium protein is an orthologous ORFan with a single homolog in M. pneumoniae. The FR result predicted that the 3D-structure of 1q8cA is similar to that of the putative regulator NusB from Mycobacterium.

Figure 1. A relatively accurate fold recognition prediction for M. genitalium’s conserved hypothetical protein MG027, 1q8cA. MG027 corresponds to an ORFan because with the exception of the close relative M. pneumoniae, it shows no sequence similarity to any other protein in the databases. Nevertheless, fold recognition is able to recognize the similarity between MG027 and a protein of known structure. The fold-recognition 3D model built without using the structural information from 1q8cA is shown on the right. The model, produced by the 3D-SHOTGUN fold-recognition method,\(^{36}\) is based on the predicted structural similarity of MG027 with the previously released structure of the transcriptor regulator NusB from Mycobacterium tuberculosis (1levyA).\(^{32}\) The sequence similarity between MG027 and 1levyA is only 13%. The prediction is confirmed by the experimental structure of 1q8cA (left). Notice that no experimental data was observed for the loop region preceding the N-terminal two helices. The overall C-alpha RMSD between the predicted model and the experimental structures is 4.9 Å, with 105 residues superimposing with an RMSD of 2.5 Å.\(^{37}\)


Additional material

An expanded Table 1, containing various links to the data is available†.

† http://bioinformatics.buffalo.edu/ORFanage/

3DORFans

References


Edited by M. Levitt

(Received 2 March 2004; received in revised form 9 June 2004; accepted 19 June 2004)