Structural bioinformatics

FCP: functional coverage of the proteome by structures

Ricard Garcia-Serna, Lulla Opatowski and Jordi Mestres*

Chernogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d’Investigació Mèdica and Universitat Pompeu Fabra, Dr Aiguader 88, 08003 Barcelona, Catalonia, Spain

Received on March 20, 2006; revised on May 10, 2006; accepted on May 11, 2006
Advance Access publication May 16, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Tools and resources for translating the remarkable growth witnessed in recent years in the number of protein structures determined experimentally into actual gain in the functional coverage of the proteome are becoming increasingly necessary. We introduce FCP, a publicly accessible web tool dedicated to analyzing the current state and trends of the population of structures within protein families. FCP offers both graphical and quantitative data on the degree of functional coverage of enzymes and nuclear receptors by existing structures, as well as on the bias observed in the distribution of structures along their respective functional classification schemes.

Availability: http://cgi.imim.es/fcp

Contact: jnestres@imim.es

1 INTRODUCTION

The complete structural resolution of the proteome would provide key information in the quest to elucidate and understand the function of proteins and their role in the processes of molecular recognition. Recent advances in high-throughput methods for protein expression and production, NMR spectroscopy and X-ray crystallography have led to an explosion in the number of protein structures determined experimentally (Sali et al., 2003). The vast majority of these structures are ultimately deposited and made publicly accessible in the Protein Data Bank (PDB), currently containing over 35 000 entries and increasing annually at an almost exponential rate (Berman et al., 2000). However, analysis of this large body of structural data reveals a natural bias towards easily tractable and therapeutically relevant proteins (Hegyi and Gerstein, 1999; Mestres, 2005), leaving large portions of the proteome devoid of representative structures. The advent of coordinated initiatives in structural genomics has started to correct this bias by selecting proteins for structure determination that specifically contribute to expanding the functional coverage of those parts of the proteome in most need of structural information (O’Toole et al., 2003; Todd et al., 2005; Xie and Bourne, 2005).

Access to information on the actual coverage of protein families by structures available in the PDB is becoming increasingly important because of its direct impact in research activities on the functional annotation of proteins (Dobson and Doig, 2005), large-scale comparative protein modeling (Pieper et al., 2006) and chemogenomic approaches to drug discovery (Bredel and Jacoby, 2004), among others. The development of FCP aims at facilitating the analysis of this structural information by providing graphical and quantitative data on the functional coverage of protein families by existing structures, as well as on the bias observed in the relative distribution of those structures among the protein members of a given family. Unfortunately, as noted above, not all the main protein families are equally represented in the PDB. As of January 2006, the family of enzymes is by far the most populated in the PDB, with 16306 entries (Laskowski, 2001, http://www.ebi.ac.uk/thornton-srv/databases/enzymes/). In contrast, 221 entries are found for nuclear receptors and only a handful is available for G protein-coupled receptors. Therefore, FCP is presently centered on enzymes and nuclear receptors.

2 DESCRIPTION

Quantifying functional coverage can only be performed if reference classification schemes exist for protein families. With current focus on enzymes and nuclear receptors, FCP has adopted the functional classification schemes recommended by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NCIUIBMB, 1992), using release 38.0 of the ENZYME Data Bank (Bairoch, 2000), and the Nuclear Receptors Nomenclature Committee (NRNC, 1999), respectively. On the basis of these reference schemes, coverage at each level of the functional classification of enzymes and nuclear receptors can be assessed by a normalized index that reflects the proportion of sublevels for which representative structures are present in the PDB relative to the total number of sublevels defined. For a given level in the classification scheme, this coverage index takes values in the range of [0,1], 1 reflecting the limiting case of all sublevels being covered by structures. However, not all covered proteins are populated with the same number of structures in the PDB. This bias has long been recognized, but few attempts have actually been made to quantify it. In this respect, FCP uses the information-theory concept of Shannon entropy to derive a normalized index that measures the variability in the distribution of structures among all sublevels within a given classification level (Mestres, 2005). Thus, for a given level in the classification scheme, the bias index takes values in the range of [0,1], with 0 reflecting the situation of a uniform distribution of structures populating all sublevels. Any deviation from uniformity is revealed by a value larger than 0.

Access to FCP from the main page is divided in three sections, namely, ‘Classification schemes’, ‘Structural and Populational filters’ and ‘Search options’. The ‘Classification schemes’ section permits direct entry to any level of the classification schemes for enzymes and nuclear receptors, providing information about current functional coverage and structural bias. To complement the
previous section, the ‘Structural and Populational filters’ section allows focus on the coverage and bias analyses on a portion of the enzyme and nuclear receptor proteome that fits certain constraints. Three structural filters are currently available that let the user select structures (1) deposited before, in or after a particular year, (2) below, at or above a certain resolution level and/or (3) coming from single or multiple sources. In addition, three populat

In the ‘Functional coverage’ section, a summary table is generated containing data on the actual number of graphical and quantitative data are provided. Among them, a specified by the filtering options selected in the main page, a variety constraints. A link to the actual data from which those distribution information on growth and coverage in function, resolution and corresponding coverage and bias indices obtained at each level. Finally, the ‘Search options’ section permits assignment of functional annotations to either a list of PDB codes or a set of proteins containing one or multiple text strings.

Once the user enters into the level of functional classification specified by the filtering options selected in the main page, a variety of graphical and quantitative data are provided. Among them, a summary table is generated containing data on the actual number of levels and sublevels populated in the PDB related to the total number defined in the classification scheme, together with the corresponding coverage and bias indices obtained at each level. In addition, several distribution graphs provide more concrete information on growth and coverage in function, resolution and organism. A link to the actual data from which those distribution graphs were created is also given. But perhaps one of the unique sources of information provided by FCP is the graph showing the historical evolution of both functional coverage and structural bias indices. As shown in Figure 1, functional coverage of enzymes increased slowly at the beginning to grow linearly in recent years at an average rate of coverage of enzymes increased slowly at the beginning to grow linearly in recent years at an average rate of structural bias above or below a certain value of their respective normalized indices. Finally, the ‘Search options’ section permits assignment of functional annotations to either a list of PDB codes or a set of proteins containing one or multiple text strings.

Once the user enters into the level of functional classification specified by the filtering options selected in the main page, a variety of graphical and quantitative data are provided. Among them, a summary table is generated containing data on the actual number of levels and sublevels populated in the PDB related to the total number defined in the classification scheme, together with the corresponding coverage and bias indices obtained at each level. In addition, several distribution graphs provide more concrete information on growth and coverage in function, resolution and organism. A link to the actual data from which those distribution graphs were created is also given. But perhaps one of the unique sources of information provided by FCP is the graph showing the historical evolution of both functional coverage and structural bias indices. As shown in Figure 1, functional coverage of enzymes increased slowly at the beginning to grow linearly in recent years at an average rate of ~100 new enzymes covered by structures per year to reach the current coverage value of 0.33. This means that there remains a total of 2612 enzymes still uncovered by structures) collecting 5141 structures (31.5% of all enzyme entries in the PDB). Excluding these 24 enzymes reduces the overall bias from 0.68 to 0.48. In spite of this, structural bias seems to be reasonably stable in recent years. This trend can be explained by the fact that the constant determination of yet more structures for enzymes already covered is being compensated by a consistent linear increase in functional coverage, resulting in retention of the overall structural bias, with a slight tendency to decrease lately.

### 3 CONCLUSION

The explosion of structural data being generated and made publicly available in the PDB necessitates tools to assess quantitatively the level of functional coverage of proteins by structures, as well as the degree of bias observed in the distribution of those structures among proteins. With FCP we provide a flexible web-based environment that facilitates analyzing, by both graphical and quantitative means, the current state and trends in the functional coverage and bias of enzyme and nuclear receptor structures at any level of their respective classification schemes, with the option to apply a variety of structural and population filters. FCP updates are synchronized with the PDB (Laskowski, 2001) and performed automatically on a quarterly basis with Python scripts. Future extensions will include analyzing functional coverage of the proteome by ligands and integrating the current functional classification schemes of proteins with gene and chemical ontologies.

### ACKNOWLEDGEMENTS

This research was supported by a grant from the Instituto de Salud Carlos III (Ministerio de Sanidad y Consumo), research project reference number 02/3051. RGS benefits from a grant from the Universitat Pompeu Fabra.

### Conflict of Interest: none declared.

### REFERENCES