Submitted for *views and challenges* section

Functional annotation strategy for protein structures

Olivia Doppelt[1,2], Fabrice Moriaud[2], Aurélie Bornot[1] & Alexandre G. de Brevern[1*]

[1] Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM UMR-S 726,
Université Denis DIDEROT - Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France
[2] MEDIT SA, 2 rue du Belvédère, 91120, Palaiseau, France.

E-mails : olivia.doppelt@medit.fr
fabrice.moriaud@medit.fr
aurelie.bornot@ebgm.jussieu.fr
alexandre.debrevern@ebgm.jussieu.fr

* Corresponding author:
mailing address: Dr. de Brevern A.G., Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM UMR-S 726, Université Denis DIDEROT - Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France
E-mail : alexandre.debrevern@ebgm.jussieu.fr
Tel: (33) 1 44 27 77 31
Fax: (33) 1 43 26 38 30

Abstract

Whole-genome sequencing projects are a major source of unknown function proteins. However, as predicting protein function from sequence remains a difficult task, research groups recently started to use 3D protein structures and structural models to bypass it. MED-SuMo compares protein surfaces analyzing the composition and spatial distribution of specific chemical groups (hydrogen bond donor, acceptor, positive, negative, aromatic, hydrophobic, guanidinium, hydroxyl, acyl and glycine). It is able to recognize proteins that have similar binding sites and thus, may perform similar functions. We present here a fine example which points out the interest of MED-SuMo approach for functional structural annotation.

Background. The number of available protein sequences has increased drastically during the last decade (472 complete genomes have been sequenced, http://www.genomesonline.org/) [1]. Still, about 40% of these sequences are characterized as "unknown function" [2,3]; they represented more than 3% [2] of the Protein DataBank (PDB) structures [4].

A majority of functional annotation methods relies on sequence similarity research, *e.g*. ProtoNet [5], or characterized sequence motifs, *e.g*. PROSITE [6]. The direct use of 3D structures or structural models to assign protein functions is an emerging field. This development is due to the increasing number of available crystallographic structures, of hypothetical proteins obtained by structural genomics consortium [7] and to new automatic crystallization methods. The first dedicated methods were directly derived from 3D local similarity methods, *i.e.* local rigid superimposition approaches. SuMo was one of the first software to use chemical groups description combined with fast graph comparison heuristic [8]. SiteEngine, developed later, had a comparable approach [9]. ProFunc is a popular web server composed of a compendium of structure-based and sequence-based methods [10].

Description. Recognition of similar binding regions on the protein surface is crucial for functional classification and for functional prediction. MED-SuMo (http://www.medit.fr/) is able to recognize proteins that have similar binding sites and thus may perform similar functions. It is an improved version of SuMo (http://sumo-pbil.ibcp.fr/ [8, 11]) with an updated source code; it is now faster and considers an increased amount of natural and synthetic ligands. Its heuristic is based on a unique representation of macromolecules using selected triplets of chemical groups which have their own geometry, regardless of the notions of main and lateral chains of amino acids. To extract similar sites, MED-SuMo transforms the binding site (or the full structure) of a query into a graph in which vertices are triplets of chemical group. Then, it is compared to binding sites extracted from the PDB which are already pre-assessed and stored in a database [11].

A major drawback in functional annotation is the difficulty to identify true “unknown function” proteins. The PDB website (http://www.rcsb.org/) associates more than 1,500 structures to an “unknown function” annotation. Nevertheless, numerous can be annotated using classical approaches (high sequence identity, structural homology, residue conservation analysis, sequence motifs research, Cleft analysis). As an illustration, 3-keto-L-gulonate 6-phosphate decarboxylase, a *lyase*, is represented by 14 proteins in the PDB. Among this family, 4 structures are classified as “unknown function”, but their functions can be found in both, the PDB and the reference paper title (*e.g.* PDB code: 1XBX). Moreover, they have significant sequence identity/similarity rate and low root mean square deviation (rmsd) with 10 other protein structures.

For our study, we have selected proteins from the “Joint Center for Structural Genomics” (JCSG, http://www.jcsg.org/); they have determined more than 350 protein structures. About half of these proteins are classified as *“Structural Genomics Unknown Function”* but most of them share sequence or fold similarity with known proteins. *Tm1012* is a hypothetical protein from *Thermotoga maritime* (PDB code: 2EWR) and cannot be associated to proteins with any known functions. Classical approach such as PSI-BLAST [12] launched on the NR database, or dedicated tool as ProFunc [5] could identify neither any related sequence nor any set of residues potentially implicated in known interaction or protein function.

As most of these methods, MED-SuMo does not give an all-or-nothing answer. The results are set out in a hit list, which are potentially interesting regions of the protein query, superimposed with corresponding similar regions of selected targets. Concerning the 2EWR query, the best hit of MED-SuMo results corresponds to the same protein crystallized by the same consortium under different experimental conditions (PDB code: 2FCL). The following hits are not directly related to the query (not superimposable, nor sharing any significant sequence identity, *i.e.* less than 20%, with 2EWR): 2CJ5, 5APR and 1OD1. 5APR and 1OD1 have a significant sequence identity rate, 38 % with a rmsd of 1.8 Å. Otherwise sequence identity rates are less than 22%. Moreover, it is not possible to superimpose any of these proteins on more than 20% of their length [13], *i.e.* these proteins are distinct.

Figure 1 shows two regions of interest for *tm1012*. The first one implicates residues 134, 135 and 138 of *tm1012* corresponding to residues 17, 18 and 31 of protein 2CJ5 (cell wall invertase inhibitor from *Nicotiana tabacum*). Figure 1a outlines the fact that these 2 proteins cannot be globally superimposed and Figure 1b displays a closer view of the local superimposition of the corresponding residues with the ligand, an acetate ion. Local rmsd is less than 0.5 Å and the two regions correspond to the same residues (YQ—L).

The second region implicates more residues. Figures 1c and 1d show the superimposition of *tm1012* and rhizopuspepsin, 5APR. Residues $S^{76}Y^{77}G^{78}D^{79}$-S^{81} of 5APR and $R^{99}L^{100}E^{101}D^{102}$-$T^{104}$ of *tm1012* are superimposed, the same residues are involved for 1OD1 (see Figure 1e). The local rmsd is quite small even if the residues are different. Interestingly only one residue (Aspartate) is common, whereas, 9 of the groups used by MED-SuMo (hydrophobic, negative, δ^+, δ^-, hydroxyl [8]) to define the ligand (statine) site of 5APR and 1OD1, are present in the binding site defined by this study for *tm1012*. Besides, analysis of the conservation of the rhizopuspepsin binding site motif shows that only 3 out of 5 residues are conserved (*Q*YG*T*-S), but 8 of the 9 groups used by

MED-SuMo are always present. These results show the interest of such an approach. They must be more deeply analyzed but gives very interesting insight for further *in silico* and *in vivo* research that could permit to functionally annotate this protein.

Conclusion. Classical sequence based approaches make possible to find more than half of the protein sequence functions. However, prediction of a protein function from its 3D structure is becoming more and more important as the worldwide structural genomics initiatives continue to solve 3D structures, many of which are unknown function proteins. We highlight the interest of MED-SuMo's heuristic, providing an application example, in which MED-SuMo is able to determine the position of a potential binding site of a hypothetical protein whereas other approaches fail. This study outlines the potential use of MED-SuMo in annotating hypothetical proteins. Further future studies will involve a large-scale analysis of many protein structures.

References.

[1] K. Liolios *et al.*, *Nucleic Acid Res*, 34: D332-4 (2006) [PMID: 16381880].
[2] I. Friedberg *et al.*, *Protein Science*, 15:1527-9 (2006) [PMID: 16731984].
[3] S. Sivashankari & P. Shanmughave, *Bioinformation*, 1:335-8 (2006) [http://www.bioinformation.net/1/80-1-2006.pdf].
[4] H.M. Berman *et al.*, *Nucleic Acid Res*, 28: 235-42 [PMID: 10592235].
[5] O. Sasson *et al. Protein Science*. 15:1557-62 (2006) [PMID: 16672244].
[6] N. Hulo *et al.*, *Nucleic Acid Res*., 34:D227-30 (2006) [PMID: 16381852].
[7] J.M. Chandonia & S.E. Brenner, *Science*. 311:347-51 [PMID: 16424331].
[8] M. Jambon *et al.*, *Proteins*, 52:137-45 (2003) [PMID: 12833538].
[9] A. Shulman-Peleg *et al.*, *J Mol Biol*, 339:607-33 (2004) [PMID: 15147845].
[10] R.A. Laskowski *et al.*, *J Mol Biol*, 351:614-26 (2005) [PMID: 16019027].
[11] M. Jambon *et al.*, *Bioinformatics*, 21:3929-30 (2006) [PMID: 16141250].
[12] S.F. Altschul *et al.*, *Nucleic Acid Res*, 25:3389-402 (1997) [PMID: 9254694].
[13] M. Tyagi *et al.*, *Nucleic Acids Res*, (2006) 34:W119-W123 [PMID: 16844973].

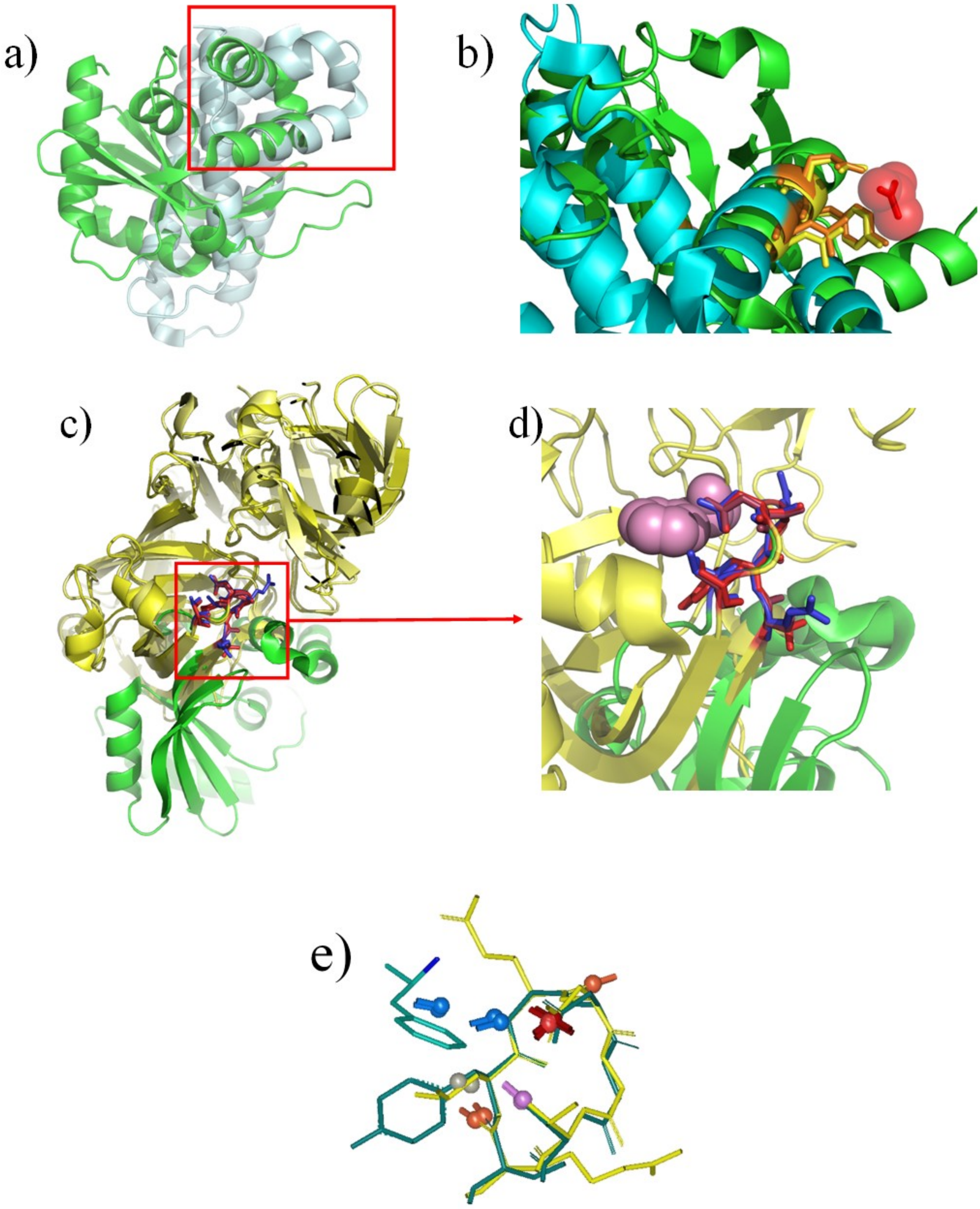

Figure 1. Examples of MED-SuMo results for hypothetical protein (*tm1012*) from *Thermotoga maritime* (in green, PDB code [4]: 2EWR, crystallized by the Joint Center for Structural Genomics, JCSG, http://www.jcsg.org/). (a) Complete superimposition of *tm1012* with the cell wall invertase inhibitor from *Nicotiana tabacum* (in blue, PDB code: 2CJ5), (b) local view of the residues superimposed by MED-SuMo, and the ligand acetate ion. (c) Superimposition of *tm1012* with rhizopuspepsin (in yellow, PDB code: 5APR), (d) with the statine ligand. (e) Superimposition of binding site of *tm1012* with Endothiapepsin (PDB code: 1OD1), the MED-SuMo groups are represented; blue: HBond donor, red: HBond acceptor, dark red: Positive, purple: hydroxyl function, light grey: hydrophobic.