Microbial Cell Factories

BioMed Central

Review

Open Access

Comparative modelling of protein structure and its impact on microbial cell factories

Nuria B Centeno, Joan Planas-Iglesias and Baldomero Oliva*

Address: Structural Bioinformatics Laboratory, Research Group on Biomedical Informatics (GRIB), IMIM/UPF. c/ Dr. Aiguader 80. 08003 Barcelona, Spain

Email: Nuria B Centeno - ncenteno@imim.es; Joan Planas-Iglesias - joan.planas@upf.edu; Baldomero Oliva* - boliva@imim.es

* Corresponding author

Published: 30 June 2005

Microbial Cell Factories 2005, **4**:20 doi:10.1186/1475-2859-4-20

Received: 03 May 2005
Accepted: 30 June 2005

This article is available from: http://www.microbialcellfactories.com/content/4/1/20

Abstract

Comparative modeling is becoming an increasingly helpful technique in microbial cell factories as the knowledge of the three-dimensional structure of a protein would be an invaluable aid to solve problems on protein production. For this reason, an introduction to comparative modeling is presented, with special emphasis on the basic concepts, opportunities and challenges of protein structure prediction. This review is intended to serve as a guide for the biologist who has no special expertise and who is not involved in the determination of protein structure. Selected applications of comparative modeling in microbial cell factories are outlined, and the role of microbial cell factories in the structural genomics initiative is discussed.

Review

Introduction

On the last two decades the development of recombinant DNA techniques has extended the use of microbial organisms to produce target proteins. The enteric bacterium *Escherichia coli* is one of the most extensively used prokaryotic organisms for genetic manipulations and for industrial production of proteins of therapeutic or commercial interest [1,2]. However, bacterial organisms often fail to produce target proteins due to problems related with protein misfolding and protein glycosilation. Yeast and fungal protein expression systems are used for the industrial production of relevant enzymes in such cases [3].

There are two main interests in the industrial production of proteins: i) Redefining the optimal properties of the target protein and ii) Avoiding problems of high-scale production. Knowledge of three-dimensional structure of the proteins may be helpful to redesign a modified protein. Computational prediction methods play an essential role to provide us with structural information of a sequence whose structure has not been experimentally determined. Homology based or comparative modeling [4] is the most detailed and accurate of all current protein structure prediction techniques [5]. Its aim is to build a three-dimensional model for a protein of unknown structure on the basis of sequence similarity to proteins of known structure [6]. Comparative modeling relies on the fact that structure is more conserved than sequence during evolution. Therefore, similar sequences exhibit nearly identical structures, and even distantly related sequences share the same fold [7,8]. Comparative modeling critically depends on the knowledge of three-dimensional structure of homologous proteins. The progress of structural genomics initiatives [9] allow to model a large amount of protein sequences. Besides, the number of unique structural folds that proteins adopt is limited {Zhang, 1997; #81; Liu, 2004 #48}. Consequently, it is likely that at least one example of most structural folds will be known, making comparative modeling applicable to most protein sequences. In term, an

Table 1: Useful servers and programs for protein comparive modeling.

PROGRAM	Server/Web adress	Reference
	Template Selection	
PSI-BLAST	http://www.ncbi.nlm.nih.gov/BLAST/	[20]
HMMER (HMM search)	http://bio.ifom-firc.it/HMMSEARCH/	[35]
TOPITS	http://www.embl-heidelberg.de/predictprotein/submit_adv.html	[17]
FUGUE	http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html	[26]
Threader	http://bioinf.cs.ucl.ac.uk/threader/	[27]
3D-PSSM	http://www.sbg.bio.ic.ac.uk/~3dpssm/	[28]
PFAM	http://www.sanger.ac.uk/Software/Pfam/	[25]
PHYLIP	http://evolution.genetics.washington.edu/phylip.html	[31]
	Target-Template alignment	
CLUSTALW	http://www.ebi.ac.uk/clustalw/	[34]
HMMER (HMM align)	http://bio.ifom-firc.it/HMMSEARCH/	[35]
STAMP	http://bioinfo.ucr.edu/pise/stamp.html	[36]
CE	http://cl.sdsc.edu	[37]
DSSP	http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html	[38]
	Model Building	
COMPOSER	http://www-cryst.bioc.cam.ac.uk	[39]
SwissModel	http://swissmodel.expasy.org/	[41]
3D-JIGSAW	http://www.bmm.icnet.uk/servers/3djigsaw/	[44]
MODELLER	http://salilab.org/modeller/	[46]
	Loop Modeling	
MODLOOP	http://alto.compbio.ucsf.edu/modloop//modloop.html	[50]
ARCHDB	http://sbi.imim.es/cgi-bin/archdb/loops.pl	[51]
Sloop	http://www-cryst.bioc.cam.ac.uk/~sloop/Browse.html	[52]
	Sidechain Modeling	
WHAT IF	http://swift.cmbi.kun.nl/whatif/	[55]
SCWRL	http://dunbrack.fccc.edu/SCWRL3.php	[56]
	Evaluation of the model	
PROCHECK	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html	[67]
PROSA II	http://www.came.sbg.ac.at/	[70]
Biotech	http://biotech.ebi.ac.uk:8400/	
	Refinement	
GROMOS	http://www.igc.ethz.ch/gromos/	[74]
CHARMM	http://www.charmm.org/	[75]
AMBER	http://amber.scripps.edu/	[76]

essential step of structural genomics is production of target proteins. Microbial cell factories play a key role in this context.

This review is intended to give a primer addressed to scientists of disciplines related to microbial cell factories who has no expertise in comparative modeling. Our goal is to provide the seeding background to understand concepts, opportunities and challenges of comparative modeling. We will describe each step in the comparative modeling process, discuss the most common errors and how to solve them, as well as outlining the applications of comparative modeling in the field of microbial cell factories.

We will emphasize the simplest and most reliable methodologies to follow up along with their range of application with a reduced number of useful programs and web servers. Many other authors have also written excellent reviews on the comparative modeling field [6,10-14].

Steps in comparative modeling

All current comparative modeling methods consist of four sequential steps: template selection, target-template alignment, model building and model evaluation. Essentially, this is an iterative procedure until a satisfactory model is obtained (Figure 1). In this process a variety of programs and web servers can be used (Table 1). Additionally, protein modeling meta-servers are emerging. They automatically implement the full process in a multi-step protocol, using simultaneously different methods [15].

Template selection

The starting point in comparative modelling is to identify protein structures related to the target sequence and then to select those that will be used as templates. Such tem-

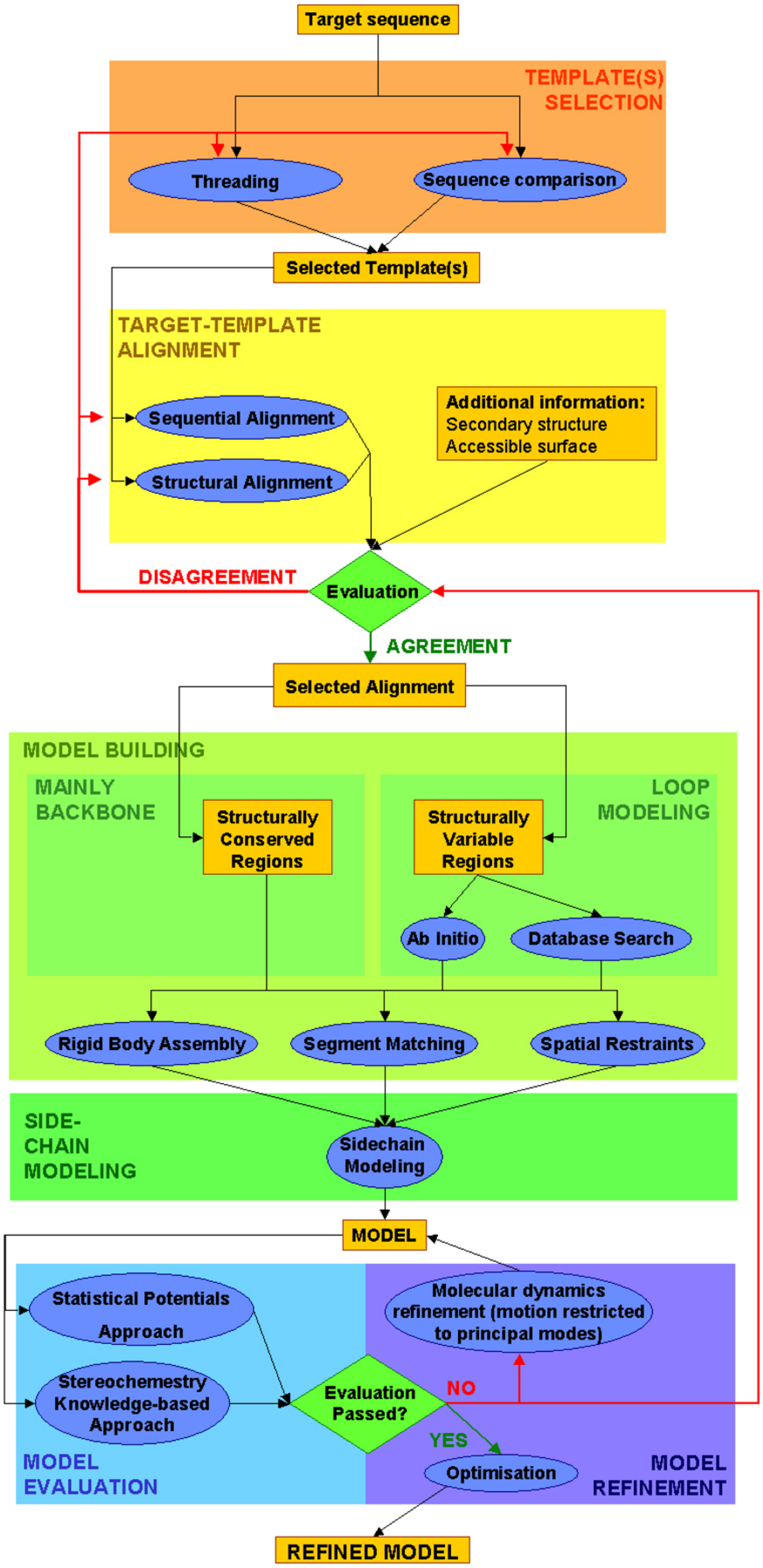

Figure 1
Flowchart of methods used for comparative modeling. Scheme of the methods used for comparative modeling, comprising template(s) selection, template-target alignment, model (backbone and loops) building, sidechain modeling, model evaluation, and model refinement steps. Programs and servers referring to these steps are listed in table 1.

plates may be found by sequence comparison methods or by sequence-structure methods also known as threading methods. Sequence comparison methods can be safely used above a certain threshold in terms of sequence identity (i.e. percentage of identical paired residues in an alignment). It has been shown that above that threshold - which is strongly dependent on sequence length, sequence homology implies structural identity [16]. Even though, below that threshold structural likeness is still possible. Some protein pairs sharing very little sequence similarity may have become similar by convergent or divergent evolution. The alignment of these proteins pairs, which define the so called "midnight zone" in sequence alignment [17], is usually addressed with threading methods. Finally, there is a range in terms of sequence identity amid the safe zone and the "midnight zone" in which the relationship between structural homology and the phylogeny is unclear: the "twilight zone" [18]. Within this range, which is usually defined between 20 and 35% sequence identity [19], additional caution must be taken on the sequence alignment.

PSI-BLAST [20], an iterative sequence comparison method, is probably the most widely used program to detect remote similarities. In more difficult scenarios, where sequence homology is not so evident, templates can be found by searching in sequence space using intermediate sequence search (ISS) methods [21-23].

Sequence comparisons can also be made through Hidden Markov Models (HMMs) [24] as implemented, for instance, in HMMER. HMMs profiles of protein domain families are available in Pfam database. These profiles can be used to automatically identify protein domain(s) within the target even if it shares weak sequence similarity with templates [25].

Threading methods have been developed to find more distant relationships. For this reason, they are the most promising choice in the absence of homologues to the target sequence. Threading methods involve performing sensitive sequence searches and characterizing sequence compatibility with the structural environments of putative templates. Features analysed by this kind of methods include secondary structure and solvent accessibility predictions as well as functional annotation. Most used methods of this kind are TOPITS [17], FUGUE [26], Threader [27] and 3D-PSSM [28]. Recent examples of the combined use of these servers and further modeling [29] prove its use. Other methods are being developed based on the analysis of protein-protein interactions to search for remote similarities [30].

Once a list of related proteins with known structure has been obtained, it is necessary to select those templates that are appropriate for the given modelling problem. The feasibility of a template can be assessed by means of its expectation value, E-value [20], which is one of the parameters in the searches outputs. As a general rule, the lower the E-value, the better the template is.

Besides, several other factors should be considered when selecting a template:

1) Quality of the experimental template structure. Because errors in templates will be passed onto the models, the better templates are the most accurate structures available. Accuracy of the templates can be assessed by the resolution and the R-factor for a crystallographic structure, or by the number of distance restraints per residue in the case of NMR structure.

2) Environment likeness. Experimental factors of interest for the target (i.e. the presence of a ligand in the structure, pH, and solvent features ...) should be found as similar as possible in the chosen templates.

3) Phylogenetic similarity. It is helpful to build a multiple alignment and a phylogenetic tree [31] of the target and templates, in order to select templates from the subfamily that is closest to the target sequence. The phylogenetic tree can be constructed by means of PHYLIP set of programs [31].

Depending on the purpose of the model, some of the factors listed above will be more important than others. For instance, resolution of the template is probably the most important factor if the reason for building the model is to design mutants of a binding site, since an accurate geometrical description is needed.

It is important to emphasize that it is not mandatory to select only one template. Actually, methods using multiple templates seems to perform better than those based on a single template [14,32], especially if the main modes extracted from them are taken into account [33]. Finally, it is noteworthy to be aware that, implicitly, choosing templates means the recognition of the target's overall fold.

Target-template alignment

Once templates have been selected, an optimal alignment between the target sequence and templates is needed to further construct a three dimensional model of the target.

From easiest to more complex, some strategies for aligning target and templates are:

1) Obtaining a multiple alignment of the templates and the target using CLUSTALW [34].

2) Aligning the query sequence to a HMM profile of the templates family built from a Pfam alignment [25] using HMMER [35].

3) Aligning the query sequence to a HMM profile of the templates built from a structural alignment using HMMER [35]. Structural alignments required for this strategy can be obtained using STAMP [36] or CE [37]; an automatic web server is available for the later one.

In our experience this third strategy, in which not only the sequence similarity but also the structural information inherent in the templates guide the alignment, make it more trustworthy.

The obtained alignments must be critically evaluated in terms of the number, length, and position of the gaps opened. Some of them can be manually refined, taking into account the secondary structure of the templates and their accessible surface, both of them calculated with DSSP program [38], in order to avoid gaps which are opened within secondary structural elements. This will be important if the alignment strategy is based only on sequence similarity. In any case, at this step, if necessary, the selection of templates may be revisited, either to search a new template to overcome a gap in a particular region or to remove redundant or inadequate templates.

Model building

Comparative model building generates an all-atom model of the sequence based on its alignment to one or more templates. It includes either sequential or simultaneous modelling of the core of the protein, loops, and side-chains.

The original comparative approach, which is still widely used, is modelling by rigid-body assembly [4]. This method constructs the model from a few regions which are obtained from dissecting related structures. In order to assemble the dissected parts, a framework is calculated by averaging of Cα atoms of structurally conserved regions in template structures. Structurally variable regions are modeled by choosing from a database of all known proteins those regions that better fit the anchor conserved regions. COMPOSER [39] is one of the programs that use this methodology in a semiautomated procedure. SwissModel is a commonly used automated web server also based in this approach [40-42].

Modeling by segment matching is another approach which relies on the approximate positions of conserved atoms in templates [27,43]. This is accomplished by breaking the target into a set of short segments, and searching in a database for matching fragments which are fitted onto an initial framework of the target structure. Database searching is based on sequence similarity, conformational similarity and compatibility with the target structure. 3D-JIGSAW is one of the successful programs that uses this approach [44]

Another approach is modeling by satisfaction of the spatial restrains obtained from the alignment [45]. Probably the most used program based on this approach is MODELLER [46]. First, this automated procedure derives many distance and dihedral angle restraints on the target sequence from its alignment with template three-dimensional structures. Next, this homology-derived restraints and energy terms ensuring proper stereochemistry are combined into a function. Finally, the model is obtained by optimizing this function in such a way that the model violates the input restraints as little as possible. Several slightly different models in agreement with the restraints can be calculated.

Any of the three methods above described produce models of similar accuracy if they are optimally applied. In the difficult cases, modeling by satisfaction of spatial restraints is perhaps the most accurate technique, since it can use many different types of information about the target sequence. In this way, available experimental data can be added as new restraints, making the model more reliable.

Loop modeling

Along with alignment, loop modeling is probably the most difficult step in comparative modelling process. Errors in loops are the dominant problem in comparative modelling when target and template share above 35% sequence identity. This is a very active area of research and it is not practical to consider all available methods (details of some of them can be found in [47-49]). In this review we will present the state-of-the art of methods that can be easily used. Furthermore, it must be pointed out that although existing methods can provide reasonably accurate models of short loop regions; modeling of long loops is still an unsolved problem [11]. Loop modeling methods can be classified in two approaches: ab initio methods and database searching or knowledge-based methods.

Ab initio loop prediction is based on a conformational search guided by a scoring or energy function -the later describing the physico-chemical properties of a protein and its environment. There are many such methods, making use of different protein representations, energy functions and optimisation procedures [11]. Among them, there is an option to use implemented in MODELLER or in a web server MODLOOP [50].

The database approach to loop prediction begins by finding segments of main chain that fit the two stems of a loop. The stems are defined as the main chain atoms that precede and follow the loop but are not part of it. The search is performed through a database of many known protein structures, not only homologues of the modeled protein. Usually, many different hits are obtained and possibly sorted according different criteria (geometric or sequence similarity). The selected segments are then superposed and annealed onto the stem regions. These initial crude models are often refined by optimisation of some energy function. Databases searching approach is more accurate and efficient if it is precedent by an structural classification of the loops present in the database. Web-servers based on structural classification [51,52] are available (see table 1). When database searching is used, it must be keep on mind that the bigger the length of the loop, the lesser the number of putative solutions that will be in the database. At the present, this fact makes this approach specially useful for loops up to 7 residues long [53].

Finally, it must be remarked that prediction of a loop conformation is hindered by two main factors: i) the exponential increase of number of possible conformations as the length of the loop grows, and ii) the conformation of a loop is influenced by the core stem regions that span the loop as well as by the structure of rest of the protein that encircles the loop. These two factors make loop modeling one of the most difficult tasks of the comparative modeling process.

Sidechain modeling

Similarly to what happens in loop modeling, sidechain conformation can be predicted from knowledge-based approaches or taking into account steric or energetic considerations [54].

Knowledge-based approaches, which are the most widely used, employ libraries of common rotamers extracted from high resolution X-ray structures. Rotamers are tried successively and scored with a variety of energy functions [13]. This approach is implemented in most automatic homology modeling procedures. Among the available software to do so it is worth mentioning: the CORALL module of WHATIF [55] and the SCWRL program [56]. Several works have probed the biological relevance of side-chain modeling, as they may imply behavioural changes in protein-protein interactions and dimerization [57-59].

Errors in comparative models

Structural models obtained by homology will have regions that resemble the true structure and regions that do not. That is, all models contain a certain amount of errors, which are more frequent as sequence identity decreases. Any stage of the comparative modeling process has its own source of errors; accordingly, they can be divided in five categories [6]:

1) Incorrect templates. This is a problem when templates share less than 25% sequence identity with the target.

2) Missalignments errors. Accuracy of the alignments is still the key limitation on the quality and usefulness of the models, being the optimal placement of gaps its limiting factor [60]. If the target and the templates have over 40% sequence identity, the alignment is almost always correct. As percentage of identity decreases, regions of local low sequence similarity appear, and alignment errors are more feasible to occur. Alignment errors increase rapidly below 30% sequence identity and become the major source of errors in this kind of models [11,14]. Target-template alignment is probably the most crucial step in comparative modelling, since any errors at this step are usually impossible to correct later [47]. Therefore, it is indeed important to devote efforts to attain the most precise alignment.

3) Structural distortions in correctly aligned regions. As sequence identity decreases, it is possible that a segment correctly aligned adopts different local structure than the target, without disruption of the overall fold. It is convenient to use multiple templates whenever they are available to overcome this problem [61].

4) Errors in regions without a template. Insertions are the most challenging regions to model, because there is not a equivalent region in the template. The complexity of the problem increases with the length of the segment. Database searching [62] or energy-based methods [63] can be applied to predict the conformation of the insertion. If there are alignment errors at stem residues or at the other environment residues, insertion modeling is not likely to result in an accurate model [49]. Therefore, the most accurate environment surrounding the insertion, the better results are obtained.

5) Errors in sidechain packing. As sequence identity decreases below 30%, there is a rapid decrease in the conservation of sidechain packing. That is, rotamers of identical residues are not conserved because the overall surroundings are changed. In addition, it must be pointed out that the correct prediction of sidechain conformation is hampered by the coupling between mainchain and sidechains and by the continuous nature of the distribution of dihedral angles [54]. This kind of error can be critical if affecting residues implicated in protein function. As we will see later, a refinement of the structure by energy minimization or molecular dynamics can sometimes surmount this problem [64].

Summarising, consequences of the errors are more serious if they are made in the initial steps of the comparative modeling process: if the selection of the template is wrong, the model based on it will be wrong; if the alignment is incorrect, local features of the model will be incorrect. Remaining errors are mainly due to incorrect description of the environment of a particular region of the structure.

Evaluation of the models

The quality of the obtained model establish the limits of the information than can be safely extracted from it. Although all structural models obtained by enclose mistakes, they become less of a problem when it is possible to detect them. Once an error is identified, it is possible to discriminate whether it affects key structural or functional regions. Accordingly, strategies to surmount errors should be taken in consideration. Therefore, an essential step in the comparative modeling process is the detection of wrongly modelled regions.

There are two different approaches to estimate errors in a structure: 1) checking the consistency of the model with experimental data of the target protein, and 2) evaluating stereochemistry and other spatial features of the model by means of methods based on statistics derived from experimentally determined protein structures.

On the first approach experimental data is used to certainly determine if particular regions of the protein are correctly modeled. Biochemical data of the most important residues regarding protein overall structure and function can be used to validate the model [65,66]. That is, they should be in close proximity in 3D space and in the correct orientation to perform their role. A consistent modeling of such residues does not ensure a good prediction; conversely, inconsistency is a important reason for concern.

One essential requisite for a model is to have a good stereochemistry. Programs used to check the stereochemistry are based in the analysis of datasets of experimentally determined protein structures. With this respect, the most widely used program is PROCHECK [67], which provide an assessment of the overall quality of the structure and highlight regions that may need further investigation.

Besides stereochemistry, there are other spatial features in the proteins, that could be used as indicators of errors in the models: packing, creation of a hydrophobic core, residue and atomic solvent accessibilities, spatial distribution of charged groups, distribution of atom-atom distances and main-chain hydrogen bonding structures [47]. This kind of information is exploited in another group of programs based on the use of energetic profiles introduced by statistical criteria [68,69]. PROSAII [70] is probably the most widely used program of this category. Although there is a concern about the theoretical validity of the energy profiles for detecting local error in models [6], this approach have been successfully applied [71,72].

It is important to note here that it is highly recommended to analyse the experimentally determined structure of templates with PROCHECK and PROSA II programs. This should allow to discriminate between errors coming from the model and errors already present in the templates.

As a final step, energy minimization and/or molecular dynamics simulations [73] of the model can be done to minimize errors detected with PROCHECK and PROSA II. The most common used programs for this purpose are GROMOS [74], CHARMM [75] and AMBER [76], which explore and evaluate the multiple possible conformations of the protein.

Performing this step is still a controversial issue [77], because the description of the physico-chemical properties of the protein and its environment is not accurate enough [11]. Even though, new evidences are suggesting that long molecular dynamics simulations with explicit solvent could overcome errors in comparative modeling [64]. Over more, strategies focusing on the appropriate sampling of biologically relevant conformations of the protein have been proved to be useful refining the model. This can be achieved by restraining the movement of specific aminoacids [78] or to particular directions in the space [33].

Comparative modeling applications in the field of microbial cell factories

On structure-function relationships

Besides other general applications of protein comparative modeling [6], there are two of them which can be of particular interest in microbial cell factories:

1) Proposing residues for site-directed mutagenesis experiments in target proteins to assess its biological function. There are many examples of how comparative modeling has been used to propose mutants, dealing with different structural features of the protein, such as electrostatic charge and surface shape [79], loop flexibility and residue accessibility [80], the protein binding or enzyme active site [81] or an enzyme alosteric site [82], among others. It is not prudent to apply comparative modeling for this purpose if the target and templates do not share at least around 30% sequence identity, since the required degree of resolution of the model will be not enough to describe the affected structural features on the target protein [6].

2) Detecting an functional important regions of a protein. The knowledge achieved in the process must allow to design proteins with altered or improved functionality. The location of a binding site can be identified by localizing clusters of charged residues [83,84] or using data of deleterious mutations [82] Biological important regions tend to be predicted better than other parts of the model [14], because amino acids in the active and binding sites are often more conserved than other structural features in a protein [85]. In addition, activity is mostly based on the physicochemical properties of residues and its spatial orientation [86]. Consequently, the degree of sequence similarity shared by the target and the templates is less restrictive for this particular application, and thus homology modeling can be applied in a wide range of scenarios, including when sequence similarity drops below 30%.

On solving protein production related problems

There are other topics in protein production processes by means of cell factories in which structural-related features play a major role. One example is protein aggregation leading to bacterial inclusion bodies, which constitute a major bottleneck in protein production [87]. Recently, it has been shown that aggregation depends on specific interactions between solvent-exposed hydrophobic stretches which adopt the form of β-sheet structures [88]. This structural knowledge provides some insight on how to solve this problem: such interactions should be specifically disrupted to avoid aggregation of β-sheets. However, full understanding of this phenomena requires also comprehending the structural details on how two or more proteins interacts. This constitutes a challenging problem known as protein-protein docking prediction [89,90]. Recent works suggest that comparative modeling can be still helpful in combination with other experimental techniques to adress this problem [91,92].

On the meeting point of comparative modeling and cell microbial factories: structural genomics

A major necessity of medium- to high-scale protein production has recently arose with the development of initiatives on structural genomics [93,94]. These initiatives, which pursue to elucidate the tree-dimensional structures of all proteins [95,96], demand optimized and further robotized protein expression systems [87]. This aim will be achieved by a focused, large-scale determination of protein structures by X-ray crystallography and NMR spectroscopy, combined efficiently with accurate protein structure modeling techniques [6].

Structural genomics, as a first step, involves ensuring that each family of proteins is represented by a known structure, avoiding unworthy efforts that will result in redundant structural information. It must be pointed out that, nowadays, there are still families of proteins which must be excluded for this kind of large-scale studies. These problematic cases include integral membrane proteins, highly disulfide-bridge proteins and large complexes [87]. All projects employ exhaustively computational methods for target selection and family exclusion [97]. For the rest of proteins, three-dimensional models can be inferred from the previously resolved family representatives. As a result, a huge amount of structural data will be available, which in turn can serve as starting point for a rational protein production design.

A complete success of the structural genomics initiative critically depends on the advances in protein production technologies. This includes new approaches in expression of targets that show challenges on protein folding [87] and also in the development of automated or semi-automated methods, robust and inexpensive for protein purification [96].

Conclusion

We have attempted to establish the capabilities and limitations of current methods of comparative modeling, as well as a general strategy to follow up in a practical case, that hopefully could serve as a guide for biologist in this field. This methods are becoming important as tools for scientists working in microbial cell factories. We have shown in this review few examples where the use of comparative modeling have been used in this area.

Comparative modeling can be safely used when target and templates share at least 30% sequence identity. Below this threshold, modeling becomes a difficult task even for experts. In any case, models must be critically evaluated to be sure that they are correct enough, devoting most of efforts to the region involved in function.

Many challenging aspects of comparative modeling are active areas of research. The state-of-the-art of the protein structure prediction strategies and methodologies is tested every two years in the CASP (Critically Assessment of techniques for protein Structure Prediction) meeting. A carefully reading of the proceedings of the meeting is probably the best way to update the progress made by the field. See supplement 6 of volume 53 of Proteins for the last report available [98].

As a final advice, it is a good policy make use of different strategies to build the model and compare them. This is always pertinent but specially as sequence identity decreases. Consistency between different models does not ensure a good prediction; however, inconsistency is a meaningful cause of concern.

With the help of structural genomics, the structure of at least one member of the most globular folds will be determined in the next years, making comparative modeling more easy. However, this is not already true for membrane proteins, which constitute a more difficult scenario [99], and more improvements in both structure determination and modeling techniques are needed.

Finally, we do believe that comparative modeling should play key role in the microbial cell factories. It will help biologists to choose which are the most interesting mutant proteins to produce, to design new proteins with a desired function, or to modify a protein to avoid production-related problems.

List of abbreviations

Target: protein to be modelled. Templates: set of proteins, homologous to the target, for which three-dimensional structure is known. Model: inferred three-dimensional structure of the target. NMR: Nuclear Magnetic Resonance. HMM: Hidden Markov Model. Structural alignment: sequence alignment based on structural similarities. Cα : Alpha-carbon; carbon atom joining the carboxyl group and the amino group in an amino acid. Restraint: as referred to in this paper, a restraint is a reduction of the conformational space of a protein on account of a prior knowledge. Main chain: sequence of atoms within a protein formed by the carboxyl group, the alpha-carbon and the amino group of each of its amino acids. Side-Chain: atoms of an amino acid not belonging to the main chain. Stem: structured boundary of a loop. Rotamer: a particular conformation of the side-chain of an amino acid regarding the position of its main chain

Authors' contributions

NBC reviewed the comparative modeling sections and updated its methods. JP reviewed the protein expression systems and the applications of comparative modeling to the microbial cell factories field. BO coordinated the design and redaction of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Authors acknowledge Fundación Ramón Areces and MCyT grant BIO2002-0369.

References

1. Baneyx F: **Recombinant protein expression in Escherichia coli.** *Curr Opin Biotechnol* 1999, **10(5):**411-421.
2. Baneyx F, Mujacic M: **Recombinant protein folding and misfolding in Escherichia coli.** *Nat Biotechnol* 2004, **22(11):**1399-1408.
3. Gerngross TU: **Advances in the production of human therapeutic proteins in yeasts and filamentous fungi.** *Nat Biotechnol* 2004, **22(11):**1409-1414.
4. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM: **Knowledge-based prediction of protein structures and the design of novel molecules.** *Nature* 1987, **326(11):**347-352.
5. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A: **Protein structure modeling for structural genomics.** *Nat Struct Biol* 2000, **7(Suppl):**986-990.
6. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A: **Comparative protein structure modeling of genes and genomes.** *Annu Rev Biophys Biomol Struct* 2000, **29:**291-325.
7. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *Embo J* 1986, **5(4):**823-826.
8. Lesk AM, Chothia C: **How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins.** *J Mol Biol* 1980, **136(3):**225-270.
9. McPherson A: **Protein crystallization in the structural genomics era.** *J Struct Funct Genomics* 2004, **5(1–2):**1-2.
10. Baker D, Sali A: **Protein structure prediction and structural genomics.** *Science* 2001, **294(5540):**93-96.
11. Fiser A, Feig M, Brooks CL 3rd, Sali A: **Evolution and physics in comparative protein structure modeling.** *Acc Chem Res* 2002, **35(6):**413-421.
12. Edwards YJ, Cottage A: **Bioinformatics methods to predict protein structure and function. A practical approach.** *Mol Biotechnol* 2003, **23(2):**139-166.
13. Krieger E, Nabuurs SB, Vriend G: **Homology modeling.** *Methods Biochem Anal* 2003, **44:**509-523.
14. Kretsinger RH, Ison RE, Hovmoller S: **Prediction of protein structure.** *Methods Enzymol* 2004, **383:**1-27.
15. Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM: **A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation.** *Proteins* 2003, **53(Suppl 6):**369-379.
16. Rost B: **Twilight zone of protein sequence alignments.** *Protein Eng* 1999, **12(2):**85-94.
17. Rost B, Schneider R, Sander C: **Protein fold recognition by prediction-based threading.** *J Mol Biol* 1997, **270(3):**471-480.
18. Doolittle RF: *Of URFs and ORFs: a primer on how to analyze derived amino acid sequences* Mill Valley, CA, USA: University Science Books; 1986.
19. Vogt G, Etzold T, Argos P: **An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited.** *J Mol Biol* 1995, **249(4):**816-831.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-3402.
21. Park J, Teichmann SA, Hubbard T, Chothia C: **Intermediate sequences increase the detection of homology between sequences.** *J Mol Biol* 1997, **273(1):**349-354.
22. Li W, Pio F, Pawlowski K, Godzik A: **Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology.** *Bioinformatics* 2000, **16(12):**1105-1110.
23. John B, Sali A: **Detection of homologous proteins by an intermediate sequence search.** *Protein Sci* 2004, **13(1):**54-62.
24. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235(5):**1501-1531.
25. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, *et al.*: **The Pfam protein families database.** *Nucleic Acids Res* 2004:D138-141.
26. Shi J, Blundell TL, Mizuguchi K: **FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties.** *J Mol Biol* 2001, **310(1):**243-257.
27. Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition.** *Nature* 1992, **358(6381):**86-89.
28. Kelley LA, MacCallum RM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J Mol Biol* 2000, **299(2):**499-520.
29. Garriga D, Diez J, Oliva B: **Modeling the helicase domain of Brome mosaic virus 1a replicase.** *J Mol Model (Online)* 2004, **10(5–6):**5-6.
30. Espadaler J, Aragues R, Eswar N, Marti-Renom MA, Querol E, Aviles FX, Sali A, Oliva B: **Detecting remotely related proteins by their interactions and sequence similarity.** *Proc Natl Acad Sci U S A* 2005, **102(20):**7151-7156.
31. Felsenstein: **Confidence-limits on phylogeneis – an approach using the bootstrap.** *Evolution* 1985, **39:**783-791.

32. Tramontano A, Leplae R, Morea V: **Analysis and assessment of comparative modeling predictions in CASP4.** *Proteins* 2001:22-38.
33. Qian B, Ortiz AR, Baker D: **Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation.** *Proc Natl Acad Sci U S A* 2004, **101(43):**15346-15351.
34. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ: **Multiple sequence alignment with Clustal X.** *Trends Biochem Sci* 1998, **23(10):**403-405.
35. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14(9):**755-763.
36. Russell RB, Barton GJ: **Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels.** *Proteins* 1992, **14(2):**309-323.
37. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein Eng* 1998, **11(9):**739-747.
38. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22(12):**2577-2637.
39. Sutcliffe MJ, Haneef I, Carney D, Blundell TL: **Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures.** *Protein Eng* 1987, **1(5):**377-384.
40. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18(15):**2714-2723.
41. Schwede T, Kopp J, Guex N, Peitsch MC: **SWISS-MODEL: An automated protein homology-modeling server.** *Nucleic Acids Res* 2003, **31(13):**3381-3385.
42. Kopp J, Schwede T: **The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models.** *Nucleic Acids Res* 2004:D230-234.
43. Unger R, Harel D, Wherland S, Sussman JL: **A 3D building blocks approach to analyzing and predicting structure of proteins.** *Proteins* 1989, **5(4):**355-373.
44. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ: **Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM.** *Proteins* 2001:39-46.
45. Havel TF, Snow ME: **A new method for building protein conformations from sequence alignments with homologues of known structure.** *J Mol Biol* 1991, **217(1):**1-7.
46. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** *J Mol Biol* 1993, **234(3):**779-815.
47. Sali A: **Modeling mutations and homologous proteins.** *Curr Opin Biotechnol* 1995, **6(4):**437-451.
48. Sanchez R, Sali A: **Advances in comparative protein-structure modelling.** *Curr Opin Struct Biol* 1997, **7(2):**206-214.
49. Fiser A, Do RK, Sali A: **Modeling of loops in protein structures.** *Protein Sci* 2000, **9(9):**1753-1773.
50. Fiser A, Sali A: **ModLoop: automated modeling of loops in protein structures.** *Bioinformatics* 2003, **19(18):**2500-2501.
51. Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJ, Oliva B: **ArchDB: automated protein loop classification as a tool for structural genomics.** *Nucleic Acids Res* 2004:D185-188.
52. Burke DF, Deane CM, Blundell TL: **Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure.** *Bioinformatics* 2000, **16(6):**513-519.
53. Fernandez-Fuentes N, Querol E, Aviles FX, Sternberg MJ, Oliva B: **Prediction of the conformation and geometry of loops in globular proteins. Testing ArchDB, a structural classification of loops.** *Proteins* 2005 in press.
54. Vasquez M: **Modeling side-chain conformation.** *Curr Opin Struct Biol* 1996, **6(2):**217-221.
55. Vriend G: **WHAT IF: a molecular modeling and drug design program.** *J Mol Graph* 1990, **8(1):**52-56.
56. Canutescu AA, Shelenkov AA, Dunbrack RL Jr: **A graph-theory algorithm for rapid protein side-chain prediction.** *Protein Sci* 2003, **12(9):**2001-2014.
57. Repiso A, Oliva B, Vives Corrons JL, Carreras J, Climent F: **Glucose phosphate isomerase deficiency: enzymatic and familial characterization of Arg346His mutation.** *Biochim Biophys Acta* 2005, **1740(3):**467-471.
58. de Atauri P, Repiso A, Oliva B, Lluis Vives-Corrons J, Climent F, Carreras J: **Characterization of the first described mutation of human red blood cell phosphoglycerate mutase.** *Biochim Biophys Acta* 2005, **1740(3):**403-410.
59. Andres AM, Soldevila M, Navarro A, Kidd KK, Oliva B, Bertranpetit J: **Positive selection in MAOA gene is human exclusive: determination of the putative amino acid change selected in the human lineage.** *Hum Genet* 2004, **115(5):**377-386.
60. Moult J, Fidelis K, Zemla A, Hubbard T: **Critical assessment of methods of protein structure prediction (CASP): round IV.** *Proteins* 2001:2-7.
61. Venclovas C: **Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment.** *Proteins* 2001:47-54.
62. van Vlijmen HW, Karplus M: **PDB-based protein loop prediction: parameters for selection and methods for optimization.** *J Mol Biol* 1997, **267(4):**975-1001.
63. Mehler EL, Periole X, Hassan SA, Weinstein H: **Key issues in the computational simulation of GPCR function: representation of loop domains.** *J Comput Aided Mol Des* 2002, **16(11):**841-853.
64. Fan H, Mark AE: **Refinement of homology-based protein structures by molecular dynamics simulation techniques.** *Protein Sci* 2004, **13(1):**211-220.
65. Carrieri A, Centeno NB, Rodrigo J, Sanz F, Carotti A: **Theoretical evidence of a salt bridge disruption as the initiating process for the alpha1d-adrenergic receptor activation: a molecular dynamics and docking study.** *Proteins* 2001, **43(4):**382-394.
66. Gutierrez-de-Teran H, Centeno NB, Pastor M, Sanz F: **Novel approaches for modeling of the A1 adenosine receptor and its agonist binding site.** *Proteins* 2004, **54(4):**705-715.
67. Laskowski RA, MacArthur MW, Moss DB, Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures.** *J Appl Crystallogr* 1993, **26:**283-291.
68. Sippl MJ: **Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.** *J Mol Biol* 1990, **213(4):**859-883.
69. Luthy R, Bowie JU, Eisenberg D: **Assessment of protein models with three-dimensional profiles.** *Nature* 1992, **356(6364):**83-85.
70. Sippl MJ: **Recognition of errors in three-dimensional structures of proteins.** *Proteins* 1993, **17(4):**355-362.
71. Zheng QC, Li ZS, Xiao JF, Sun M, Zhang Y, Sun CC: **Homology modeling and PAPS ligand (cofactor) binding study of bovine phenol sulfotransferase.** *J Mol Model (Online)* 2005.
72. Aloy P, Mas JM, Marti-Renom MA, Querol E, Aviles FX, Oliva B: **Refinement of modelled structures by knowledge-based energy profiles and secondary structure prediction: application to the human procarboxypeptidase A2.** *J Comput Aided Mol Des* 2000, **14(1):**83-92.
73. Hansson T, Oostenbrink C, van Gunsteren W: **Molecular dynamics simulations.** *Curr Opin Struct Biol* 2002, **12(2):**190-196.
74. Scott WRP, Hunenberger PH, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF: **The GROMOS biomiolecular simulation program package.** *J Phys Chem A* 1999, **103:**3596-3607.
75. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: **CHARMM – A program for macromolecular energy, minimization, and dynamics calculations.** *J Comput Chem* 1983, **16(2):**513-519.
76. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, Debolt S, Ferguson D, Seib el G, Collman P: **AMBER, a package of computer-programs for applying molecular mechanics,<normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules.** *Comput Phys Commun* 1995, **91:**1-41.
77. Tramontano A, Morea V: **Assessment of homology-based predictions in CASP5.** *Proteins* 2003, **53(Suppl 6):**352-368.
78. Flohil JA, Vriend G, Berendsen HJ: **Completion and refinement of 3-D homology models with restricted molecular dynamics: application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis.** *Proteins* 2002, **48(4):**593-604.
79. Iverson GM, Reddel S, Victoria EJ, Cockerill KA, Wang YX, Marti-Renom MA, Sali A, Marquis DM, Krilis SA, Linnik MD: **Use of single point mutations in domain I of beta 2-glycoprotein I to**

determine fine antigenic specificity of antiphospholipid autoantibodies. *J Immunol* 2002, **169(12):**7097-7103.

80. Feliu JX, Benito A, Oliva B, Aviles FX, Villaverde A: **Conformational flexibility in a highly mobile protein loop of foot-and-mouth disease virus: distinct structural requirements for integrin and antibody binding.** *J Mol Biol* 1998, **283(2):**331-338.
81. Sathyanarayanan PV, Siems WF, Jones JP, Poovaiah BW: **Calcium-stimulated autophosphorylation site of plant chimeric calcium/calmodulin-dependent protein kinase.** *J Biol Chem* 2001, **276(35):**32940-32947.
82. Gloyn AL, Odili S, Zelent D, Buettger C, Castleden HA, Steele AM, Stride A, Shiota C, Magnuson MA, Lorini R, *et al.*: **Insights into the structure and regulation of glucokinase from a novel mutation (V62M), which causes maturity-onset diabetes of the young.** *J Biol Chem* 2005, **280(14):**14105-14113.
83. Matsui E, Abe J, Yokoyama H, Matsui I: **Aromatic residues located close to the active center are essential for the catalytic reaction of flap endonuclease-1 from hyperthermophilic archaeon Pyrococcus horikoshii.** *J Biol Chem* 2004, **279(16):**16687-16696.
84. Ishino T, Pasut G, Scibek J, Chaiken I: **Kinetic interaction analysis of human interleukin 5 receptor alpha mutants reveals a unique binding topology and charge distribution for cytokine recognition.** *J Biol Chem* 2004, **279(10):**9547-9556.
85. Valdar WS, Thornton JM: **Conservation helps to identify biologically relevant crystal contacts.** *J Mol Biol* 2001, **313(2):**399-416.
86. Villa-Freixa J, Bonet J, Khan AK, Johnston M: **Evaluating the relationship between residue stability and enzyme active site preorganization in protein regulated enzymes.** *Theor Chem Acc* 2005 in press.
87. Yokoyama S: **Protein expression systems for structural genomics and proteomics.** *Curr Opin Chem Biol* 2003, **7(1):**39-43.
88. Carrio M, Gonzalez-Montalban N, Vera A, Villaverde A, Ventura S: **Amylod-like properties of bacterial inclusion bodies.** *J Mol Biol* 2005, **347(5):**1025-1037.
89. Wodak SJ, Mendez R: **Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications.** *Curr Opin Struct Biol* 2004, **14(2):**242-249.
90. Schneidman-Duhovny D, Nussinov R, Wolfson HJ: **Predicting molecular interactions in silico: II. Protein-protein and protein-drug docking.** *Curr Med Chem* 2004, **11(1):**91-107.
91. Seri M, Savino M, Bordo D, Cusano R, Rocca B, Meloni I, Di Bari F, Koivisto PA, Bolognesi M, Ghiggeri GM, *et al.*: **Epstein syndrome: another renal disorder with mutations in the nonmuscle myosin heavy chain 9 gene.** *Hum Genet* 2002, **110(2):**182-186.
92. Almstedt K, Lundqvist M, Carlsson J, Karlsson M, Persson B, Jonsson BH, Carlsson U, Hammarstrom P: **Unfolding a folding disease: folding, misfolding and aggregation of the marble brain syndrome-associated mutant H107Y of human carbonic anhydrase II.** *J Mol Biol* 2004, **342(2):**619-633.
93. Kim SH: **Shining a light on structural genomics.** *Nat Struct Biol* 1998, **5(Suppl):**643-645.
94. Goldsmith-Fischman S, Honig B: **Structural genomics: computational methods for structure analysis.** *Protein Sci* 2003, **12(9):**1813-1821.
95. Montelione GT, Anderson S: **Structural genomics: keystone for a Human Proteome Project.** *Nat Struct Biol* 1999, **6(1):**11-12.
96. Edwards AM, Arrowsmith CH, Christendat D, Dharamsi A, Friesen JD, Greenblatt JF, Vedadi M: **Protein production: feeding the crystallographers and NMR spectroscopists.** *Nat Struct Biol* 2000, **7(Suppl):**970-972.
97. Brenner SE: **Target selection for structural genomics.** *Nat Struct Biol* 2000, **7(Suppl):**967-969.
98. **CASP5. Proceedings of the 5th Meeting on the Critical Assessment of Techniques for Protein Structure Prediction. 1–5 December Asilomar, California, USA.** *Proteins* 2002, **53(Suppl 6):**333-595.
99. Becker OM, Shacham S, Marantz Y, Noiman S: **Modeling the 3D structure of GPCRs: advances and application to drug discovery.** *Curr Opin Drug Discov Devel* 2003, **6(3):**353-361.