structure-function relationships in this novel molecule.

Since the identification of *H. pylori* almost 20 years ago, basic studies of pathogenesis have identified numerous areas in which new concepts have been uncovered, including discovery of autolysis as a mechanism for surface-association of cytoplasmic proteins, host-mediated phosphorylation of bacterial proteins delivered to human gastric epithelial cells[14] and a novel family of penicillin-binding proteins[15,16]. The report by Ha *et al.*[1] adds another concept — short peptide segments added to existing structures may allow assembly of higher-order oligomers with properties (in this case, ability to resist inactivation by acid) not seen in the original structures. The conclusions of this study thus take us one step further toward understanding the ability of *H. pylori* to resist acid and perhaps the development of new treatments for eradicating *H. pylori* infection.

*Bruce E. Dunn is in the Department of Pathology, Medical College of Wisconsin, and Pathology and Laboratory Medicine Service, Milwaukee Department of Veterans Affairs Medical Center, Milwaukee, Wisconsin 53295, USA. Markus G. Grütter is at Biochemisches Institut, Universität Zürich, Zürich, Switzerland. Correspondence should be addressed to B.E.D. email: Bruce.Dunn@med.va.gov*

1. Ha, N. *et al. Nature Struct. Biol.* **8**, 505–509 (2001).
2. Jabri, E., Carr, M. B., Hausinger, R.P. & Karplus, P.A. *Science* **268,** 998–1004 (1995).
3. Benini, S. *et al. Structure Fold. Des.* **7**, 205–216 (1999).
4. Dunn, B. E. *et al. Infect. Immun.* **65**, 1181–1188 (1997).
5. Phadnis, S. H. *et al. Infect. Immun.* **64**, 905–912 (1996).
6. Krishnamurthy, P. *et al. Infect. Immun.* **66,** 5060–5066 (1998).
7. Weeks, D., Eskandari,S., Scott, D.R. & Sachs, G. *Science* **287,** 482-485 (2000).
8. Scott, D.R. *et al. Gastroenterology.* **114**, 58-70 (1998).
9. Miederer, S. E. & P. Grubel. *Digestive Diseases Sci.* **41,** 944–949 (1996).
10. Marshall, B. J. *et al. Am. J. Gastroenterol* **82,** 200 (1987).
11. Stingl, K. *et al. Infect. Immun.* **69**, 1178–1180 (2001).
12. Austin, J.W., Doig, P., Stewart, M. & Trust, T.J. *J. Bacteriol.* **174**, 7470-7473 (1992).
13. Chu, S., Tanaka, S., Kaunitz, J.D. & Montrose, M.H. *J. Clin. Invest.* **103**, 605–612 (1999).
14. Stein, M., Rappuoli, R. & Covacci, A. *Proc. Natl. Acad. Sci. USA* **97**, 1263–1268 (2000).
15. Krishnamurthy, P. *et al. J. Bacteriol.* **181**, 5107–5110 (1999).
16. Mittl, P.R., Luthy, L., Hunziker, P. & Grütter, M.G. *J. Biol. Chem.* **275**, 17693–17699 (2000).

# Target practice

Andrej Sali

**The scope of structural genomics has recently been estimated by simulation of several target selection strategies based on the currently known protein sequence families. Useful characterization of most protein sequences will be possible by protein structure modeling, if structures of ~16,000 carefully selected protein domains are determined experimentally. In the absence of globally coordinated target selection, three times as many structures may be required.**

Structural genomics is a comprehensive effort toward characterizing the structures of all proteins[1–11]. The first essential step in structural genomics is the selection of target protein sequences for experimental structure determination such that all the remaining proteins are related to at least one known structure at a useful level of similarity (Fig. 1). On pages 559–566 of this issue of *Nature Structural Biology*, Vitkup *et al.*[12] describe the scope of structural genomics. The number of targets is estimated from similarities among the sequences in the Pfam database of 2,000 family alignments of related protein domains[13]. To relate 90% of the domain sequences in Pfam to a known structure with >30% sequence identity, two structures per Pfam family are needed. The Pfam domain families cover only a quarter of the domains in several representative genomes. In practice, inefficiencies in target selection are estimated to increase the number of targets by approximately a factor of three relative to the optimal target selection. Thus, the scope of structural genomics corresponds up to 50,000 targets, which is well within reach of the nascent global structural genomics effort[14].

*A priori*, structural genomics targets can be qualified by two criteria. First, the targets are likely to correspond to individual domains rather than multidomain proteins. The reason is that the structure of a single domain is usually easier to determine by X-ray crystallography or NMR spectroscopy than that of a more flexible multi-domain protein. Second, domains that are not amenable to structure determination are excluded from consideration. Such domains may include membrane spanning domains, domains containing long regions of unusual amino acid residue composition of low complexity, large flexible domains, domains that require ligands for stability and variants resulting from post-translational modifications and alternative splicing.

Target selection is intimately tied to the specific aim of structural genomics. For example, if the aim is to map distant evolutionary relationships between all related domains[15], only a relatively low-density sampling of the protein space is required. In contrast, the inability of protein structure modeling to reliably predict functional differences between homologs led others to include close homologs on the target list (for example, 70% sequence identity); however, in that case the scope is limited to a single genome so that the project is still feasible[9]. Many additional target selection strategies of different groups involved in structural genomics initiatives are reviewed in ref. 15. For example, target lists may correspond to the representatives of all fold families[16,17], functional families[7], all proteins from a genome[3] or all unusual uncharacterized soluble proteins in a small genome[18]. Domain families and domain sequences may be prioritized by relevance and feasibility criteria, such as currently perceived medical importance and the absence of transmembrane spanning helices, respectively. The target lists of the individual research groups are usually limited to a certain type of protein

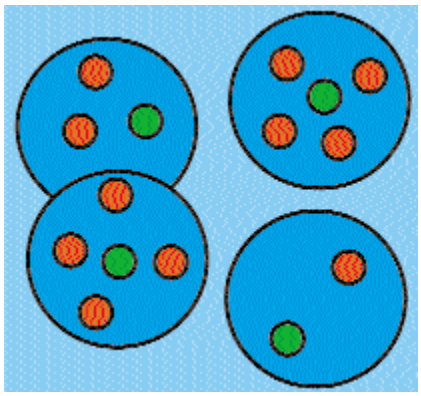

**Fig. 1** Selecting protein domain targets for the global structural genomics effort. Once the domains appropriate for structure determination in all known proteins are defined, the only remaining step in target selection is to cluster these domains into the smallest possible number of groups using appropriate criteria. The targets for structural genomics are any one of the domains in groups without any structurally defined members. The criteria for the clustering of domains determine the number of targets; they depend on the modeling ability and the available resources. The groups of domains might correspond to the different fold types, superfamilies of domains originating from the common ancestors, families of domains that share at least 30% sequence identity, or families of domains with the 'same' functions. The final result of structural genomics will be a map of the protein sequence-structure space (the light blue rectangle) that relates all similar domains (red and green dots) and provides structures for at least one member (green dots) of each domain group (dark blue circles), as well as 3D models for all of the remaining sequences (red dots).

(for example, cancer-related proteins) or to a subset of all protein sequences (for example, a genome) to make the size of the individual projects reasonable. In contrast to individual groups who can afford to focus on relatively small parts of the protein space, the target selection of the global structural genomics effort must cover all protein sequences that are amenable to structure determination.

For the purposes of target selection, it is convenient to take a model-centric view of structural genomics: structural genomics aims to produce useful comparative models for most protein sequences[12,19]. This view is justified because the first step in many structure-based annotations is the calculation of a comparative model[20], although there are trivial cases where modeling is not needed and difficult cases where modeling is not helpful. To obtain a reasonable level of accuracy, the models must be based on alignments with few errors. Such alignments can usually be obtained when the sequence identity between the modeled sequence and at least one known structure is higher than 30% (ref. 20). Thus, structural genomics should determine protein structures so that most sequences in the genome databases match at least one structure with an overall sequence identity of more than 30% (refs 12, 19).

Vitkup *et al.*[12] first estimate the number of structural genomics targets for a well defined set of 2,000 protein domain families in the Pfam 4.4 database. The targets are selected by a 'greedy' coverage algorithm. This simple algorithm picks a target iteratively by maximizing the number of domain sequences that can be modeled based on at least 30% sequence identity to the selected target structure. The number of targets required to cover all of the 260,000 domain sequences in Pfam is 17,000 (13,000 if the membrane spanning domains are excluded). Above 30% sequence identity, the number of targets increases by 10,000 per 10 percentage points of sequence identity.

As described below, Vitkup *et al.*[12] quantify substantial reductions in the number of targets that result from improving modeling techniques and from relaxing the completeness requirement. They also address the negative impacts of failure in structure determination and deviations from the optimal target selection strategy.

The number of required targets would be reduced by a factor of two if the modeling techniques were improved so that the accuracy of comparative models based on 20% sequence identity equaled the current accuracy at 30% sequence identity[12]. To achieve this aim, improvements in all aspects of comparative modeling are required, including fold assignment, sequence-structure alignment, and modeling of insertions, core segments, and sidechains[20].

A substantial reduction in the number of targets can also be achieved if outliers in large families and some small families are initially ignored. For example, when the coverage requirement is relaxed from 100% to 90% of all sequences in Pfam, only 4,000 targets (2 per family) instead of 17,000 targets (8 per family) are required[12].

On the downside, it might be expected that the efficiency of structural genomics is decreased significantly by the low success rate of structure determination — that is, 10–20% for randomly picked protein sequences[9]. However, the corresponding decrease in the coverage of domain sequences by structural genomics is only 10% (ref. 12). The reason is that large families provide many alternative targets, most of which are satisfactory because they allow modeling of many of the remaining family members. This result supports the class-directed approach to structure determination, which is at the core of structural genomics[1].

The efficiency of structural genomics is also reduced when the individual research groups apply different target selection criteria[12]. They may not all rigorously use the 30% sequence identity cutoff and may impose additional filters, such as the genome of origin and the biological significance of the target. As a consequence, the 'selection' of targets for the global structural genomics effort does not minimize the number of targets required for structural characterization of most protein sequences. The target selection efficiency in practice is expected to correspond to that of selecting targets randomly, but only if they have <30% sequence identity to an already determined structure. In that case, three times as many targets would be required as with the optimal 'greedy' algorithm. This result provides a strong incentive for global coordination of target lists. Steps in this direction include the web sites of the individual research groups mandated by the National Institutes of Health (NIH) in North America[14], web sites with comprehensive target lists (*http://presage.berkeley.edu*, *http://www.structuralgenomics.org*), and tools such as PartsList, a web based system for dynamically ranking domain folds based on more than 180 attributes[21].

The final step in estimating the scope of structural genomics is to cautiously extrapolate from the number of targets needed for the current Pfam domain families to the number of targets needed for all domain families[12]. It is necessary to assume that the modeling density in Pfam applies to all domain families, including the currently unknown ones. Since only about a quarter of all residues in the coding regions of several representative genomes match one of the 2,000 Pfam families, the total number of protein domains is estimated to be ~8,000, which is consistent with some other estimates[22]. Because 12,000 targets are required to cover 90% of sequences in the current Pfam database when using target selection without global coordination, the scope of a comprehensive structural genomics effort is up to 50,000 targets (including the membrane spanning domains). In other words, if the structures of 50,000 target domains are determined by experiment, it should be possible to model ~90% of all sequences based on at least 30% sequence identity.

In comparison, the fraction of domains that can currently be modeled based on at least 30% sequence identity to a known structure is only ~10% (ref. 12). Thus, the currently known structures do not significantly reduce the scope of structural genomics if at least 30% sequence identity is required for modeling.

At present, structural biologists are producing ~500 protein structures qualifying as structural genomics targets per year[12]. In a few years, the global structural genomics effort is likely to exceed this number several fold. Thus, in my view it is conceivable that structures of 70% of all protein domains within boundaries of structural genomics will be structurally characterized in less than five years. As a result, application of the powerful principles of structural biology to most biological problems is imminent.

**Acknowledgments**
A.S. is grateful to S.K. Burley, J. Kuriyan, T. Gaasterland and other members of the New York Structural Genomics Research Consortium for many discussions about structural genomics, and to H.M. Moss and N. Eswar for comments on the manuscript. A.S. is an Irma T. Hirschl Trust Career Scientist. Support from The Merck Genome Research Institute, Mathers Foundation, and NIH is also acknowledged.

*Andrej Sali is in the Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA. email: sali@rockefeller.edu*

1. Terwilliger, T.C. *et al. Protein Sci.* 7, 1851–1856 (1998).
2. Sali, A. *Nature Struct. Biol.* 5, 1029–1032 (1998).
3. Zarembinski, T.I. *et al. Proc. Natl. Acad. Sci. USA* **95**, 15189–15193 (1998).
4. Montelione, G.T. & Anderson, S. *Nature Struct. Biol.* **6**, 11–12 (1999).
5. Teichmann, S.A., Chothia, C. & Gerstein, M. *Curr. Opin. Struct. Biol.* **9**, 390–399 (1999).
6. Burley, S.K. *et al. Nature Genet.* **23**, 151–157 (1999).
7. Cort, J.R., Koonin, E.V., Bash, P.A. & Kennedy, M.A. *Nucleic Acids Res.* **27**, 4018–4027 (1999).
8. Brenner, S.E. & Levitt, M. *Protein Sci.* **9**, 197–200 (2000).
9. Christendat, D. *et al. Nature Struct. Biol.* **7**, 903–909 (2000).
10. Heinemann, U. *Nature Struct. Biol.* **7**, 940–942 (2000).
11. Yokoyama, S. *Nature Struct. Biol.* **7**, 943–945 (2000).
12. Vitkup, D., Malamud, E., Moult, J. & Sander, C. *Nature Struct. Biol.* **8**, 559–566 (2001).
13. Bateman, A. *et al. Nucleic Acids Res.* **27**, 263–266 (2000).
14. *Nature Struct. Biol.* **7**, 927–994 (2000).
15. Brenner, S.E. *Nature Struct. Biol.* **7**, 967–969 (2000).
16. Mallick, P., Goodwill, K.E., Fitz-Gibbons, S., Miller, J.H. & Eisenberg, D. *Proc. Natl. Acad. Sci. USA* **97**, 2450–2455 (2000).
17. Portugalyi, E. & Linial, M. *Proc. Natl. Acad. Sci. USA* **97**, 5161–5116 (2000).
18. Balasubramanian, S., Schneider, T., Gerstein, M. & Regan, L. *Nucleic Acids Res.* **28**, 3075–3082 (2000).
19. Sánchez, R. *et al. Nature Struct. Biol.* **7**, 986–990 (2000).
20. Martí-Renom, M.A. *et al. Rev. Biophys. Biomolec. Struct.* **29**, 291–325 (2000).
21. Qian, J. *et al. Nucleic Acids. Res.* **29**, 1750–1764 (2001).
22. Wolf, Y.I., Grishin, N.V. & Koonin, E.V. *J. Mol. Biol.* **299**, 897–905 (2000).

# β-catenin and its multiple partners: promiscuity explained

Lawrence Shapiro

**β-catenin functions both in cadherin-mediated cell adhesion and the Wnt signaling pathway. In these roles, β-catenin interacts with a multitude of protein partners including cadherins, α-catenin, axin, APC, and Tcf family transcription factors. Recent crystal structures show how β-catenin can achieve this remarkably diverse functionality.**

Some proteins seem to have too many different jobs — so much so that it can be difficult to understand how they do it all. β-catenin is a champion jack-of-all-trades: it functions both in cadherin-based cell adhesion and also as a central component of the developmentally important Wnt signaling pathway, dysfunction of which is a common cause of human cancers. In these different roles, β-catenin has a multitude of specific binding partners. As structural biologists, we can be skeptical of such diverse functions until we see at atomic level details of their specificity and possible mechanisms. Recent papers by Graham *et al.*[1] and Huber and Weis[2] in *Cell* clearly show how β-catenin recognizes two of its most important targets: the transcription factor Tcf, which is the ultimate activator of downstream Wnt signaling, and the cytoplasmic region of E-cadherin. These structures show that β-catenin recognizes at least a subset of its binding partners as elongated peptides, through a set of 'quasi-independent' subsites that may be regulated independently of one another. This suggests how β-catenin can play such a central role in its different functions.

## The many roles of β-catenin

In its cell adhesion role, β-catenin associates with the cytoplasmic domain of cadherin cell–cell adhesion proteins, such as E-cadherin (Fig. 1). Through a distinct binding site, β-catenin also binds to the unrelated protein α-catenin, which in turn binds to filamentous actin. Thus, β-catenin provides the linkage between transmembrane cadherin adhesion proteins and the cytoskeleton. This linkage to the cytoskeleton is crucial to the cell adhesion function of cadherins because cadherins cannot mediate adhesion in cells that lack either α- or β-catenin.

But that's just part of the story. In its signaling role, β-catenin binds specifically to adenomatous polypsosis coli (APC), the product of the infamous colon cancer gene whose main function is to degrade β-catenin. Wnt signaling is an important developmental pathway, regulated by Wnt family growth factors (wingless is the *Drosophila* Wnt). In the absence of a Wnt signal, phosphorylation of β-catenin by glycogen synthase kinase-3β (GSK-3β) targets β-catenin for degradation. This phosphorylation event occurs in a multiprotein complex that contains APC and axin, both of which bind β-catenin. In the presence of a Wnt signal, GSK-3β phosphorylation is inhibited, and the cytoplasmic pool of β-catenin is stabilized. This free β-catenin moves to the nucleus where it binds to Tcf transcription factors. The β-catenin–Tcf complex activates transcription of downstream genes that effect the Wnt signaling program.

So, β-catenin binds to (at least) the following partners: cadherins, α-catenin, axin, APC, and the Tcf transcription factors, which share little apparent sequence