ELSEVIER

Dynamic contact maps of protein structures

Erik L. L. Sonnhammer and John C. Wootton

Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

The two-dimensional contact map of interresidue distances is a visual analysis technique for protein structures. We present two standalone software tools designed to be used in combination to increase the versatility of this simple yet powerful technique. First, the program Structer calculates contact maps from three-dimensional molecular structural data. The contact map matrix can then be viewed in the graphical matrix-visualization program Dotter. Instead of using a predefined distance cutoff, we exploit Dotter's dynamic rendering control, allowing interactive exploration at varying distance cutoffs after calculating the matrix once. Structer can use a number of distance measures, can incorporate multiple chains in one contact map, and allows masking of user-defined residue sets. It works either directly with PDB files, or can use the MMDB network API for reading structures. © *1998 by Elsevier Science Inc.*

INTRODUCTION

To comprehend and analyze protein structures efficiently, it is necessary to reduce the complexity of the picture given by all atom positions. This can be done at the three-dimensional level, where, for instance, only the backbone may be shown, with secondary structure elements marked as easily recognizable objects. Alternatively, the topology of the secondary structure elements may be further condensed into a two-dimensional diagram.[1]

A different way of viewing any three-dimensional structure is based on the *contact map*, which, in its simplest form, is a matrix of all pairwise distances within the molecule. The axes of the contact map are linear in the sequence of the protein chain, which makes it attractive for relating structural features to the sequence. Many protein structure analysis packages therefore include a contact map module.[2]

We present here a contact map calculation program, Structer, which is coupled to a matrix visualization program, Dotter.[3] Dotter was originally designed not for viewing contact maps, but for sequence similarity dot plots. For such dot plots, the calculation part is built in. However, viewing a matrix of bytes is a generic task, and since Dotter can read matrices from file it can be used to view any type of matrix. The main advantage of using Dotter is that the rendering of a pixel matrix can be changed dynamically on the screen, obviating the need to recalculate the matrix with a different distance cutoff. This provides a higher degree of interactivity between the analyst and the data.

Structer has a number of additional features, such as masking of selected residue types and application of different interresidue distance measures. The distances can also be converted to neighbor ranks. Structer can also write the contact map matrix in ASCII format for viewing in programs other than Dotter.

METHODS

Structer and Dotter were written in C. Dotter uses the ACEDB graphics library, which at present is supported for Unix workstations running X-Windows. The script Dotstructer, for running both Structer and Dotter as a single command, was written in Gawk.

To export a contact map to Dotter, Structer prepares two files: a sequence file and a pixel-matrix file. The sequence file is in FASTA format, and the pixel matrix is in Dotter format 1, which is a simple format wherein the pixels are passed as a stream of bytes. The pixel bytes are not raw gray tones but a "score" that is later converted to a gray tone depending on Dotter's current setting. The pixel bytes are preceded by a header with the following fields: fileformat (unsigned char, 1 byte), zoomfactor (int, 4 bytes), horizontal_len (int, 4 bytes), and vertical_len (int, 4 bytes). The fileformat and zoomfactor were set to 1, and horizontal_len and vertical_len were set to the length of the sequence. All int fields must be stored with the most significant byte first.

The most important thing to keep in mind is that for tech-

The Color Plate for this article is on page 33.

Address reprint requests to: Erik L. L. Sonnhammer, Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, Building 38A, Room 8N805, National Institutes of Health, Bethesda, Maryland 20894.

Received 27 March 1998; Accepted 15 April 1998.

655 Avenue of the Americas, New York, NY 10010

1093-3263/98/$19.00
PII S1093-3263(98)00009-6

nical reasons, horizontal_len and vertical_len must be the smallest multiple of 4 greater or equal to the actual sequence length. If the sequence is, e.g., 197 long, horizontal_len and vertical_len must be set to 200, and the pixel matrix must contain the corresponding amount of pixels. For example, if the matrix was made from a sequence of length 197, the pixel map must thus contain 200 × 200 pixels. For more details, see http://www.sanger.ac.uk/Software/Dotter.

The MMDB network API[4] is part of the NCBI toolkit.[5]

RESULTS

Structer and Dotter

The program Structer calculates a contact map from the atomic coordinates of a molecular structure, which is either read from a PDB structure file or via the network by calling the MMDB network API. By default, a contact map is calculated on the basis of all chains included in the PDB entry, but it can also be limited to one or a few chains. Since all contact maps calculated from metric distances are symmetrical across the main diagonal, Structer normally calculates only half of the matrix for efficiency reasons, but mirrors it to the other half. The copying to the mirror image can optionally be turned off.

After calculating the contact map, Structer can prepare a matrix of pixel values to be read into Dotter for visual inspection. By separating the tasks of contact map calculation and visualization, we achieve a higher degree of flexibility than in the fully integrated case. Updates of the two programs can proceed independently, which facilitates maintenance, and if several distinct matrix calculation routines were built into Dotter, it would become too unwieldy. The user normally wants to use the two programs in one action, however; to this end a simple shell script Dotstructer is available, which first calls Structer and then calls Dotter with the Structer output files.

Structer also contains a number of functions that are not used for Dotter; for instance, it can output the matrix in a straightforward ASCII format for use in another viewer, or it can output a table of coordinates with PDB and MMDB residue numbering, and MMDB domain numbering. It can also convert the sequence in a PDB file to FASTA format. For such functions it is more practical to have a separate, nongraphical program.

Figure 1A and B and Color Plate 1[6] show a sample view of a Dotter contact map produced by Structer. The middle domain of diphtheria toxin, a transmembrane helix bundle, produces characteristically thick diagonals, while the C-terminal domain, a β barrel, produces much thinner diagonals. The N-terminal domain is a mixed α–β domain and yields a more complex picture. In this example (PDB:1ddt), the coordinates of residues 188–199 are unknown; by default Structer suppresses such residues. They may optionally be included if MMDB is

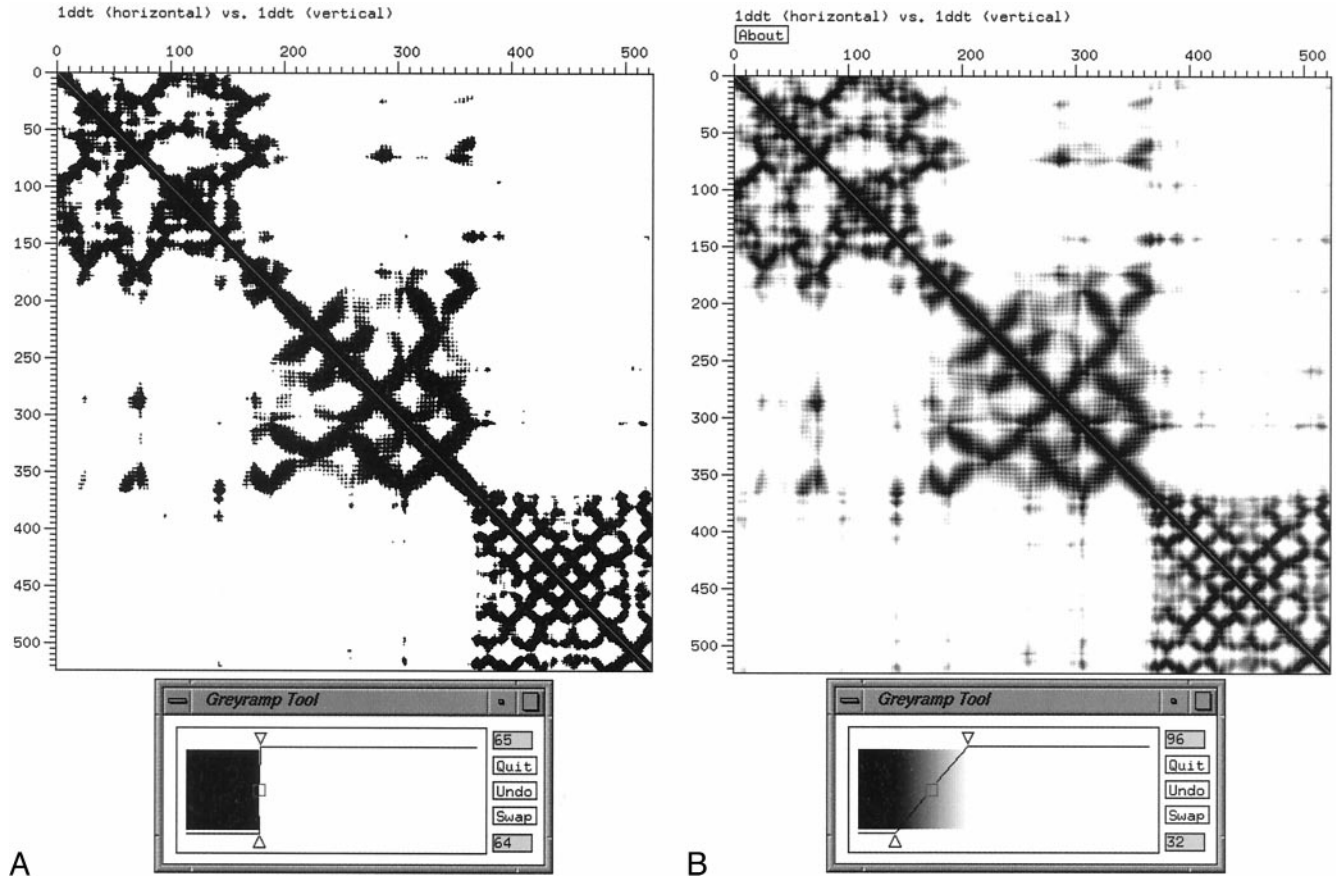

Figure 1. Structer C_α contact map of diphtheria toxin (PDB:1ddt) shown in Dotter. (A) The Greyramp tool is set so that pairwise distances less than 20 Å are black, and all greater distances are white. (B) Same matrix, but with the Greyramp tool set so that distances between 10 and 30 Å are shown in gray tones proportional to the distance.

used to fetch the structure. The part of the contact map containing residues with unknown coordinates will then be set to the maximum distance (highest pixel value).

Dotter's Greyramp tool

The function of the Greyramp tool requires some explanation. It is a general-purpose tool, which is part of the ACEDB graphics library, for dynamically changing the rendering of pixels in windows that contain so-called pixel maps.

As seen in Figures 1A and B, the "ramp" in the Greyramp tool has two cutoffs; one at the top and one at the bottom. The current cutoff values are shown in text boxes at the right. The principle of the ramp is that pixel values between the two cutoffs are rendered in a gray tone linearly proportional to its position between the two cutoffs. For instance, a pixel value half-way in between will be 50% gray. Pixel values outside the two cutoffs are either 100% black or white. Values outside the bottom cutoff are rendered black and values outside the top cutoff are white. This allows focusing of the gray range on a narrow range of pixel values. The coloring can be inverted by setting the top cutoff to a value below the lower cutoff (pressing the Swap button effects this). The inverted setting is default in Dotter since it is more intuitive for sequence similarity dot plots, where a higher score is more relevant. For contact maps, however, the normal setting is more intuitive, since here black means that the residue pair is in contact. Hence the Dotter image should be inverted when used with Structer by clicking once on the Swap button.

Figures 1A and 1B show the contact map generated by Structer as viewed by Dotter using two different distance renderings. Changing the rendering of the matrix is effected by dragging the triangular or square controls on the Greyramp tool, or by typing in pixel value cutoffs in the text boxes. Usually, the top and bottom cutoffs are kept at a fixed distance from each other while both are changed synchronously by dragging the middle square. Setting the two cutoffs only one value apart is sometimes preferable, for maximum contrast, as shown in Figure 1A. It is often useful to slide the cutoffs in this high-contrast mode to produce a "melt-through" animation. For static pictures with more depth, a wider range between the cutoffs is usually preferable (Figure 1B).

Structer scales the contact map to fit the 0–255 pixel value scale. This is done either by providing a fixed scale, e.g., 10 per angstrom, in which case the pixel values are easy to interpret, or by setting the scale so that the maximum distance in the contact map equals a pixel value of 255. The latter is the default in Structer; the thus-calculated scale factor is reported to the user. In Figure 1, the scale was 3.2 Å per pixel value.

Distance measures

Structer can measure the distance between two residues by three different methods: the distance between the C_α atoms, the average interatomic distance, or the shortest interatomic distance. The C_α distance is the default method since it is much faster to calculate; the contact maps from the other methods differ mainly in the details. All atoms are used for the average and shortest interatomic distance calculations, except between adjacent residues in the chain, where only side-chain atoms are considered. For glycine, the C_α is counted as the side chain.

The contact map can further be transformed to a *contact-rank map*, in which the distances are transformed to rank numbers, i.e., the closest residue is assigned rank 1, the next rank 2, etc. This method was proposed by Karlin et al.[7] for measuring residue associations, and has been used for several structural analyses (e.g., Brocchieri and Karlin[8] and Karlin and Zhu[9]). Note that the rank of one residue relative to another is not reciprocal, so the contact map matrix is no longer symmetrical across the main diagonal. Contact-rank maps thus necessitate display of the entire contact map matrix. Residues that protrude from the core tend the have high ranks with most other residues. These residues thus form vertical lines in the contact map; see Figure 2. Conversely, residues in the core tend to have low to medium ranks with most other residues. The contact-rank map can thus be useful for discerning core and loop regions.

Contact maps of multiple chains

To analyze the points of interaction between different protein chains, it is useful to display a contact map of multiple chains. By default, Structer produces a contact map of all protein chains in a PDB entry. The user can also instruct Structer to limit the contact map to a selection of the available chains. An example with multiple chains, human class II MHC with a bound peptide, is shown in Figure 3. The start of each chain is marked with a label (PDB and chain identifiers), and can

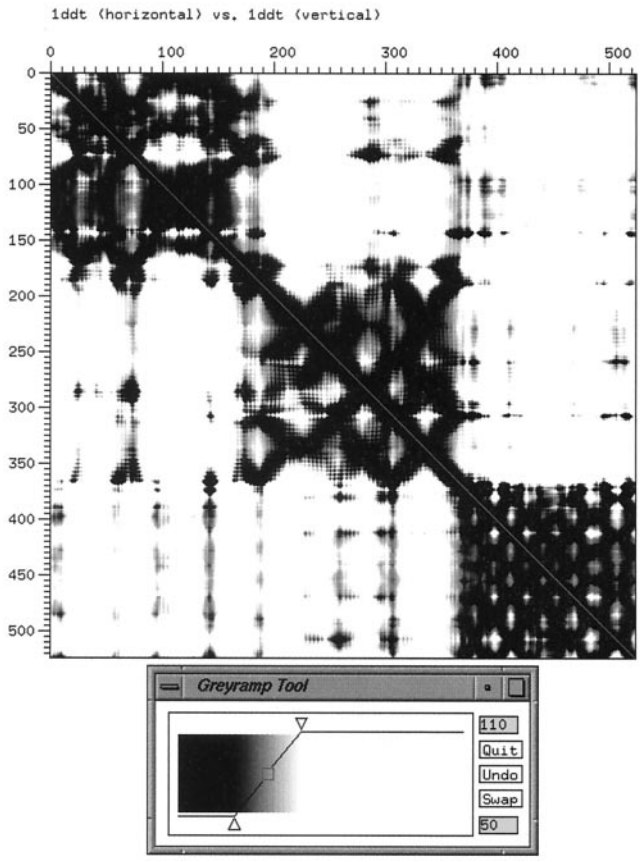

Figure 2. Structer/Dotter contact-rank map of the protein in Figure 1, showing protruding elements in the two N-terminal domains.

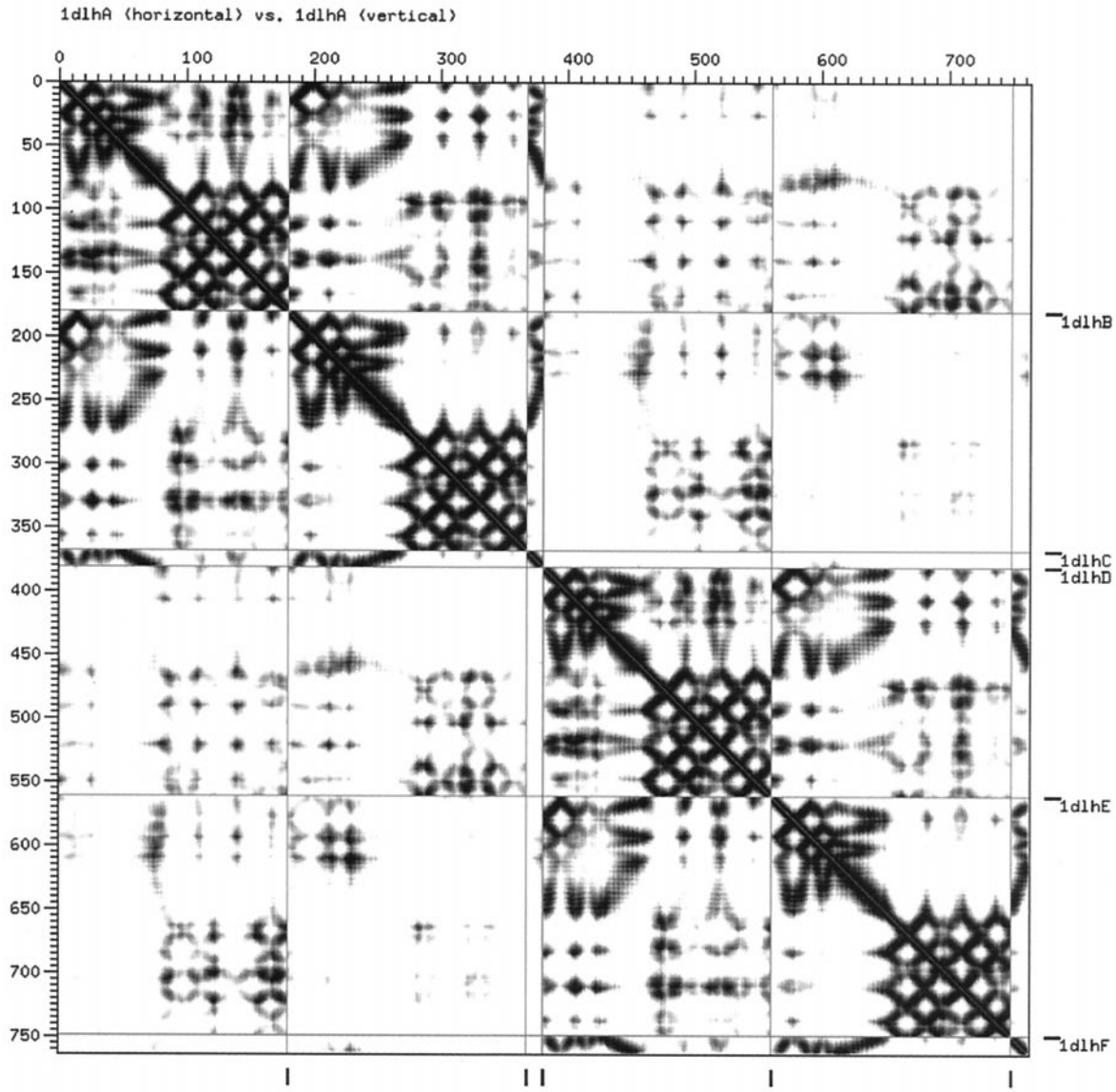

Figure 3. Structer/Dotter contact map of human class II MHC protein (PDB:1dlh) complexed with a virus peptide (1dlhC), showing the interchain contacts.

optionally be highlighted with a line by selecting "Draw lines at segment ends" in Dotter's "Feature Series Selection Tool."

Selecting subsets of residues

For examining which residues make contact in a given contact map, two facilities are provided in Structer/Dotter. First, Dotter has a crosshair that can be positioned over the pixel in question, and the residue pair can be read out at the center of Dotter's Alignment tool. Second, Structer can limit the contact map to only certain residue pairs. For instance, if a contact map of only hydrophic residues is wanted, all pairs consisting of AFILMPVW can be selected. Pairs that contain other residues are masked out by being assigned a pixel value corresponding to the maximum distance. The selection of residue subsets proceeds by specifying a set of allowed residues for either residue of a pair. The wildcard "*" can be used to allow any residue on one side. For instance, "AFILMPVW-*" shows all pairs that have at least one hydrophobic residue. An example of this is shown in Figure 4.

Availability

Structer and Dotstructer are available via anonymous FTP from ncbi.nlm.nih.gov in /pub/esr/structer/, Dotter in /pub/esr/dotter/, and the MMDB API in /pub/mmdb/mmdbapi/.

DISCUSSION

The residue–residue contact map of a protein structure can be used for many analytical purposes. It can be a powerful tool for detecting spatial features in structures if the user is accustomed to interpreting contact map patterns. We hope that by providing a modern and flexible system for contact map analysis, this type of analysis will become more widespread. The dynamic distance threshold control of Structer/Dotter should facilitate

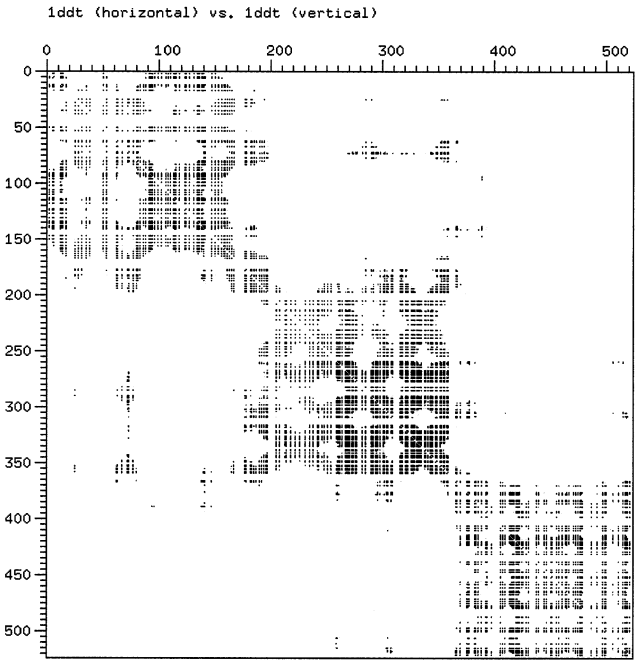

Figure 4. Structer/Dotter contact map of the protein in Figure 1, showing only hydrophobic residue pairs. This illustrates that the two N-terminal domains have a hydrophobic core localized toward the C terminus of the domain, while the C-terminal domain has an evenly distributed core.

exploration significantly, and the ability to display multiple chains in one contact map should be useful for interchain contact analysis.

At present, all contact maps are produced at zoom level 1, i.e., one residue pair always corresponds to 1 pixel. For large proteins (>1000 residues) that produce contact maps larger than the screen, the viewed area is controlled with scrollbars.

In this article we have described methods for analyzing only known protein structures. We would like to stress that Dotter would also be an excellent tool for exploring predicted contact or interaction maps,[10–12] especially since these tend to contain high levels of noise that could be filtered out with Dotter's Greyramp tool.

ACKNOWLEDGMENT

We thank Chris Hogue for assistance with linking the MMDB API.

REFERENCES

1 Flores, T.P., Moss, D.S., and Thornton, J.M. An algorithm for automatically generating protein topology cartoons. *Protein Eng.* 1994, **7,** 31–37

2 Shindyalov, I.N., and Bourne, P.E. WPDB—PC Windows-based interrogation of macromolecular structure. *J. Appl. Crystallogr.* 1995, **28,** 847–852

3 Sonnhammer, E.L.L., and Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 1995, **167,** GC1–10

4 Hogue, C.W. Cn3D: A new generation of three-dimensional molecular structure viewer. *Trends Biochem. Sci.* 1997, **22,** 314–316

5 Ostell, J.M. The NCBI software tools. In: *DNA and Protein Sequence Analysis: A Practical Approach* (Bishop, M.J., and Rawlings, C.J. eds.), IRL Press, Oxford, 1997, pp. 31–43

6 Sayle, R.A., and Milner-White, E.J. RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci.* 1995, **20,** 374

7 Karlin, S., Zuker, M., and Brocchieri, L. Measuring residue associations in protein structures. Possible implications for protein folding. *J. Mol. Biol.* 1994, **239,** 227–248

8 Brocchieri, L., and Karlin, S. How are close residues of protein structures distributed in primary sequence? *Proc. Natl. Acad. Sci. U.S.A.* 1995, **92,** 12136–12140

9 Karlin, S., and Zhu, Z.Y. Characterizations of diverse residue clusters in protein three-dimensional structures. *Proc. Natl. Acad. Sci. U.S.A.* 1996, **93,** 8344–8349

10 Selbig, J. Contact pattern-induced pair potentials for protein fold recognition. *Protein Eng.* 1995, **8,** 339–351

11 Hubbard, T.J., and Park, J. Fold recognition and ab initio structure predictions using hidden Markov models and beta-strand pair potentials. *Proteins* 1995, **23,** 398–402

12 Mirny, L., and Domany, E. Protein fold recognition and dynamics in the space of contact maps. *Proteins* 1996, **26,** 391–410