Sukhjeet Batth

Professor Douglas Brutlag

Biochemistry 118Q

1 June 2004

Tertiary Protein Structure Prediction with Profile Analysis: A Case Study

The ability to predict protein structure from amino acid sequences alone would be invaluable to scientists everywhere. A pharmaceutical company working on creating a new life-saving drug, for example, could analyze the sequence of a protein that plays an integral role in the disease with a computer and find its structure in seconds. This sort of bioinformatics method is far more time and cost efficient than the long, tedious, and costly analyses of structural biologists, which include x-ray crystallography and nuclear magnetic resonance. The present paper does not provide an overview of current protein structure prediction methods. However, it goes through, step-by-step, how to predict a specific type of suspect structure by utilizing a set of sequences of known structure. This method is explained by using a case study of tertiary structure prediction with Profile Analysis.

Proteins play an integral part in nearly all biological processes. Various types of proteins include structural, transportation, hormonal, contractile movement, antibody, and enzyme. These macromolecules are known as polypeptides because they are formed when multiple amino acids join to form peptide bonds. Proteins can be viewed from four different levels of structure: primary, secondary, tertiary, and quaternary. The primary structure is the sequence of amino acids. The secondary structure consists of alpha-helices and beta-sheets formed from hydrogen bonding. Tertiary structure encompasses protein folding and stabilization with the formation of

disulfide and hydrogen bonds. Finally, the quaternary structure is the interaction of two or more polypeptide chains.

Alpha-helices are a very common component of proteins and have several characteristics. This type of structure forms from hydrogen bonding between the carboxyl and amino groups of different amino acids. The mean amino acids per turn of a helix are 3.6 or one per 100 degrees of rotation (Branden and Tooze). The length of an alpha-helix, on average, is approximately ten residues or three turns (Branden and Tooze).

The discovery of alpha-helices as models for the local folding of polypeptide chains opened up the doors to much research of this new secondary structure (Pauling, Corey, Branson 1951). Several models of alpha-helix packing were suggested soon after. Francis Crick proposed the earliest alpha-helix packing model (Crick 1953). Crick's "knobs into holes" model suggested that helices pack by fitting amino acids of one helix (knobs) into spaces formed by amino acids of another helix (holes). He also created a method to study helices at a planar level through superimposed helical lattices. Alexander Efimov proposed another model that attempted to relate the interaxial angle of a packed helix pair with preferred rotational states of the side-chains (Efimov 1979). He proposed that the side-chains affected packing due to their unique chemistry.

Approximately twenty-five years ago, Chothia *et al*. proposed a now widely accepted model for helix-to-helix packing of proteins in the formation of globular protein molecules (Chothia, Levitt, Richardson 1977). The conceptual basis for the model relates the ridges and grooves formed by amino acids as they spiral to form a

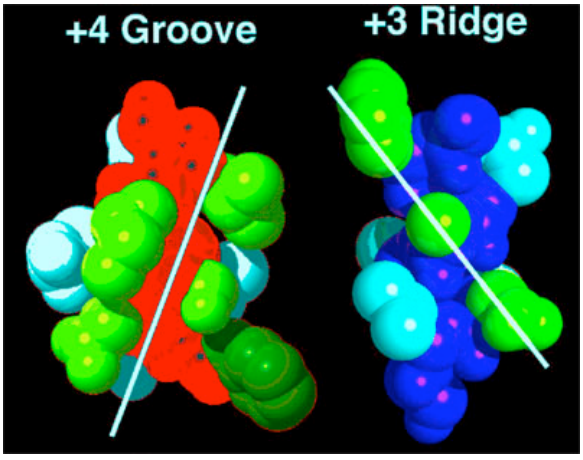

Figure 1 – A +4 groove and a +3 ridge

helix (Figure 1). Packing is the fitting of the ridge of one helix into the groove of another ("ridges into grooves"). There are four types of helices that vary in the number of residues in between points on a straight line drawn along a ridge. The two types of packing helices that will be discussed are +3 and +4. A +3 helix has an intrahelical packing at every third position along the ridge (*e.g.* i, i+3, i+6 where i is the initial contact residue) whereas a +4 helix has a ridge with intrahelical packing at every fourth position (*e.g.* i, i+4, i+8).

There are three main packing classes that arise from differences in the shapes of helix surfaces. These differences can be seen by looking at arrangements of the amino acids as they come together when helices pack, also known as packing diamonds. The three main packing classes are 1-4, 4-3, and 4-4. Each number in a pair represents the number of amino acids in between points along the ridges of an alpha-helix. For example, in 4-4 packing there are two packing helices that both have packed amino acids along the ridge at every fourth position.

The 4-3 packing class, shown in Figure 2, has an expected interaxial angle of approximately +20 degrees, which is the tightest interaxial angle of all packing classes, indicating that it is the closest to ideal packing. Parallel packing is ideal because it utilizes the maximum surface area. This type of packing class (4-3) has specific packing diamonds. A packing diamond for this class is the network of amino acids formed by the contacts at the i, i+3, i+4, and i+7 positions (solid, yellow lines in Figure 2). Most

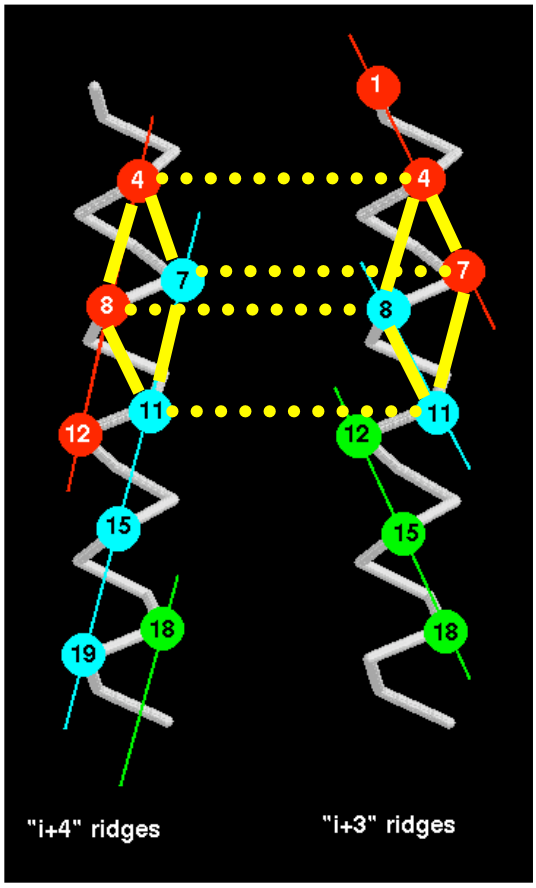

Figure 2 – Packing diamonds and contacts depicted by solid and dotted lines

packing is anti-parallel, in which an i from one helix contacts an i+7 from the other helix in the packed pair.

Profile Analysis is a method that can be used to predict desired characteristics in protein sequences (Gribskov, Luthy, Eisenberg 1990). This method includes the use of profiles for scoring sequences of similar structure. A profile consists of a table of numerical values that reflect the probabilities of amino acids being located at given positions for expressing desired characteristics (Table 1). These numerical values for probabilities are derived directly from sequences in a sequence file used to create the profile along with a structure correlated scoring matrix. The profile can then score a sequence based on similarities between the tested sequence and the sequences used to create the profile.

Sample Profile

+3 Helix Profile (i to i+11)
3heli11.prf

Pos	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
i	-49	-50	-45	-66	-21	54	-44	-5	196	-27	73	17	-33	-127	-23	46	-87	60	186	40	4	20	-23
i+1	19	4	6	34	19	-2	-48	-6	-19	107	-1	-50	-25	-21	1	110	5	2	-18	49	7	-19	10
i+2	183	-50	-86	-21	145	22	-107	-25	-1	108	101	67	-79	-20	50	80	-69	-98	-19	-14	7	-101	97
i+3	297	-57	-73	-7	221	-6	-165	-51	-31	160	232	136	-108	-10	96	23	-161	-176	-33	-35	10	-184	158
i+4	98	-54	-76	-40	86	17	-121	-35	132	83	144	74	-68	-89	29	21	-119	-24	117	-13	8	-38	56
i+5	-2	98	-36	83	9	-31	30	-7	8	37	27	-11	114	-48	0	34	-40	-25	15	-47	7	-35	4
i+6	16	35	-45	-1	9	43	-22	-6	62	28	32	52	72	-75	-9	0	-58	15	29	-25	5	-22	-1
i+7	78	-18	-68	-17	85	39	-111	-15	95	43	190	76	-18	-94	37	0	-133	-5	101	-29	10	-68	60
i+8	44	-54	-58	-49	60	54	-85	-19	94	17	85	85	-58	-89	21	2	-85	28	127	-15	5	-24	40
i+9	64	51	-45	73	68	-54	-42	-4	-10	37	61	38	30	-42	128	0	-56	-45	18	-48	11	-63	98
i+10	-38	-28	10	-38	-27	60	-5	9	11	-37	-36	10	-19	-32	-24	1	37	37	10	0	-1	98	-26
i+11	23	5	-16	10	27	-20	9	2	-5	7	40	18	0	-17	54	-6	-30	-22	-2	-20	2	-29	40

Table 1 – A sample profile with scores for amino acids at specific positions

Presently, there is a lack of knowledge that links the primary structure to tertiary, folded structure. The primary structure, the sequence of amino acids, is largely responsible for determining all higher levels of protein structure and function. The twenty distinct monomers of proteins, amino acids, are chemically diverse; there are groups of hydrophobic, hydrophilic, acidic, and basic amino acids. These distinct characteristics of the different amino acids make them appropriate for different functions. A method like Profile Analysis exploits this fact and

can be used to take sequences with known structures, create a weighted scoring matrix, and test sequences that are suspected to have structural similarities to the sequences of known structure used.

Chothia *et al.* previously studied this type of packing and identified different sets of proteins with segments that packed in this way (Chothia, Levitt, Richardson 1981). These helices included contact residues that had been determined by x-ray crystallographic data (Table 2).

Protein Name	PDB Name	Helix Name		Helix Pair	PDB Range	Chothia Range
Bovine Carboxy-peptidase A	5CPA	CPA	124	1	14-28	14-29
			182	2	72-88	72-90
Lysozyme Bacteriophage T4	7LZM	LZM	82	4	82-90	82-91
			93	5	93-106	93-106
			93b	5	93-106	93-106
			142	10	143-155	142-156
			114	7	115-123	114-123
			126	8	126-134	126-134
Subtilisin BPN' Precursor	1ST2	SBT	210	4	n/a	103-118
			239	5	n/a	132-146
Thermolysin	8TLN	TLN	159	4 (Chothia: 3)	160-180	159-181
			261	10 (Chothia: 5)	260-274	261-274
Tobacco Mosaic Virus Coat Protein	2TMV	TMV	38	3 (Chothia: 2)	37-52	38-51
			76	4 (Chothia: 3)	73-86	76-87
			76	4 (Chothia: 3)	73-86	76-87
			114	5 (Chothia: 4)	111-135	114-133

Table 2 – Lists of helices that 4-3 pack according to Chothia *et al.*

After creating packing diagrams depicting which residues were in contact and of which packing diamond they were a part, multiple sequence files aligned by the initial contact residues were created with the use of Genedoc (Nicholas, Deerfield 1997). A number of sequence files were created to test different ranges of the helices: i to i+7, i to i+11, i-2 to i+9, i-2 to i+11, i-4 to i+9, and i-4 to i+11. Separate sequences files were created from helix ranges specified by PDB and Chothia *et al.* for both +3 and +4 helices.

Profilemake, part of the Genetics Computing Group suite of programs, was used to make profiles from the files of known helices and a structure-correlated scoring matrix. In addition, leave-one-out profiles were made by omitting one sequence from the set of sequences used to create the profiles. The Gribskov Method is the basis for the calculation of these scores (Gribskov, Luthy, Eisenberg 1990). This method includes creating weighted scores for residues at specific positions in protein sequences. A sequence can be tested with a profile, which creates a sequence score that is the sum of weighted scores for residues at specific positions in the sequence. The next step was to score the known helices with the profiles. Profile-SS, a version of the Profile Search and Scan program, was used because it included both negative and positive scores in the outputs (Ropelewski, Nicholas, Gribskov 1987). This allowed for determination of the range around the first contact in packing that should be used for prediction.

After different ranges were tested in this way, a leave-one-out analysis with the sequence files was done to determine the ranges of accuracy for various predictions. A protein with an unknown structure, cholecystokinin-58 (CCK-58), was used for this. If the z-score for the prediction of the first packing contact in CCK-58 was within the minimum and maximum z-scores for all of the predictions for the known helices, then it was considered reliable prediction. Furthermore, the predictions should pick positions conserved throughout various CCK-58 sequences if this packing is significant. CCK-58 is a major regulatory peptide found in the cerebral cortex and small intestine. Some major functions of this protein are gallbladder contraction and pancreatic enzyme release. Although there is limited information about the structure of this protein, there is some data suggesting it has a unique structure and that it has helices (Keire *et al.* 2002). All of these characteristics made CCK-58 an interesting candidate for this test.

The results indicated that the predictions for CCK-58 were reliable and a position was predicted that was conserved throughout the sequences from various species. It is known that the biological and immunological activity of CCK-58 is different than its close relatives, CCK-8 and CCK-33, which share the same binding regions. This suggests that this sort of packing may actually occur in CCK-58. Furthermore, it predicts exactly where this type of packing is likely to happen.

This case study took Profile Analysis and showed, step-by-step, how it can be applied to predict tertiary structure. Profile Analysis was typically used to identify homologous sequences based on the pattern of amino acid properties selected for during protein evolution. However, when applied in this way, it can identify a specific type of suspect tertiary structure. This is a good example of how we do not necessarily have to find something new to get closer to the goal of successful protein structure prediction, but rather sometimes look at what we already know in a different way.

Works Cited

Branden, Carl, and John Tooze. Introduction to Protein Structure. New York: Garland Publishing, 1991.

Chothia, Cyrus, Michael Levitt, and Douglas Richardson. "Helix to Helix Packing in Proteins." Journal of Molecular Biology 145 (1981): 215-250.

---. "Structure of proteins: Packing of alpha-helices and pleated sheets." Proceedings of the National Academy of Sciences, USA. 74.10 (1977): 4130-4134.

Crick, Francis H.C. "The Packing of alpha-Helices: Simple Coiled-Coils." Acta Crystallographica 6 (1953): 689-697.

Efimov, Alexander V. "Packing of alpha-helices in globular proteins: layer structure of globin hydrophobic cores." Journal of Molecular Biology 236 (1979): 1356-1368.

Gribskov, Michael, Roland Luthy, and David Eisenberg. "Profile Analysis." Methods in Enzymology 183 (1990): 146-159.

Keire, David A., Travis E. Solomon, and Joseph R. Reeve, Jr. "NMR evidence for different conformations of the bioactive region of rat CCK-8 and CCK-58." Biochemical and Biophysical Research Communications 293.3 (2002): 1014-1020.

Nicholas, Karl B., Hugh B. Nicholas, Jr., and David W. Deerfield, II. "GeneDoc: Analysis and Visualization of Genetic Variation." EMBNEW News 4 (1997): 14.

Pauling, Linus, Robert B. Corey, and H. R. Branson. "The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chains." Proceedings of the National Academy of Sciences, USA 37.4 (1951): 205-211.

Ropelewski, Alexander J., Hugh B. Nicholas, and Michael Gribskov. Profile-SS: Optimal Profile-Sequence Alignment Program. Vers. 1.8. Pittsburgh Supercomputing Center, 1987.