Eric Bennett
Computational Molecular Biology
Final Project
6/6/02

Utilization of primary and secondary structure elements to predict a protein's propensity to form amyloids.

Introduction

Protein aggregation and amyloid formation has long been tied to pathological disorders (1). The cell expends a large amount of its energy towards properly folding its proteins, regulating their destruction, and preventing non-productive associations leading to aggregate and inclusion formation (2). Breakdowns in the protein folding pathway have been shown to decrease the cell's ability to clear cellular waste leading to a buildup of cellular trash. This trash has been linked to numerous neurodegenerative diseases and much research effort has been aimed at elucidating the misfolding pathway. Many of these conformational diseases have been shown to be genetically encoded in which a particular mutated protein leads to improper folding pathways. Huntingtin's and Parkinson's disease arise from mutations in single proteins giving rise to an aggregated state that either escapes or overwhelms the cellular clearance mechanisms (3-5). One of the most well characterized conformational diseases is Alzheimer's disease. A hallmark of the cytopathology is the accumulation of amyloid plaques in neuropils and neurofibrillar tangles (6). All of these diseases point to a common feature. These aggregation prone proteins are normally soluble with a variety of three-dimensional structures that are stable until some mutation or cellular event triggers a conformational change leading to amyloid formation (7-9). Amyloid fibrils have an extremely characteristic structure in which they are built up from β-strands perpendicular and parallel to the main fiber axis (10-11). It appears that different proteins can adopt this structure arguing that the amyloid conformation might be merely

an alternative folding conformation amenable to all proteins regardless of their amino acid makeup (12,20). However, it also must be true that sequence elements dictate the propensity to form aggregates as huntingtin proteins with a glutamine expansion of 25 remains soluble whereas CAG trinucleotide expansion beyond 40 results in aggregation formation (13-15). Also, the kinetics of aggregation in Huntingtin's models has also been tied to the length of the trinucleotide repeats.

Despite the ability of aggregation prone proteins to form common amyloid structures, the structures of the soluble proteins as well as their individual cellular functions are extremely divergent (16). How can proteins so distantly related all follow a similar misfolding pathway? It is possible that all of the amyloid forming proteins share a common structural motif that can be gleaning from analysis of their primary amino acid sequence. However, attempts to identify a common sequence characteristic among these proteins have failed. Studies of the amyloid β (Aβ) peptide found in amyloid plaques of Alzheimer's patients have shown that a conformational switch from a α-helix to a β-sheet precipitates the aggregation event (17,22). This α/β trigger has also been shown to proceed the misfolding event in prion diseases (18,19,21). This commonality might serve as the avenue for which computational techniques can identify possible amyloid forming proteins from sequence analysis and secondary/tertiary structure predictions.

Prediction of secondary structure based solely on the information within a protein sequence is highly useful in a number of different respects. Methods for predicting secondary structure have improved substantially in the last decade through the use of dynamic programming methods and evolutionary information gleaned from proteins in the same structural family (23-25). To date, the best prediction algorithms reach a predictive capability of 76%,

meaning that 76% of all residues were correctly predicted as helix, strand, or other (26). This places prediction techniques at the level of resolution equal to the best efforts of Fourier-transform infrared (FT-IR) and circular dichroism (CD) structural techniques (27). Recent increases in the accuracy of prediction stems from three sources. First, and most significantly, the sheer increase in the number to sequences and alignments in Genbank along with the concurrent rise in the number of three-dimensional structures available have increased sensitivity by allowing the dynamic programming methods to be trained on a more complete set of structures (28,29). It is easy to image that increasing the known diversity of sequences that give rise to common structural elements will greatly aid prediction capabilities. Second, the prediction methods themselves have been refined combining multiple neural networks to achieve increased accuracy (30-33). Lastly, the ability to construct multiple sequence alignments that include more distant homologs using PSI-BLAST has contributed to recent advances in prediction accuracy (34,35). However, it can also be said that accuracy has really only increased 7-10% since the first iteration of PHD in the early 90s. Given smaller proteins with 100 amino acids, these advances are marginal at best. Although, small increases in prediction accuracy can be extremely beneficial when analyzing larger more complex proteins especially if predictions are made on a genome-wide scale. The holy grail of computational structural prediction is *ab initio* 3-D structure prediction using only sequence information to construct the model. However, combinatorial methods using known structural homologs in the protein data bank as well as predicted secondary structure have led to the best overall tertiary structure predictions.

Results

These recent advances in prediction capabilities might aid in predicting structural motifs underlying amyloid formation. A first-glance analysis at the sequences of the known

aggregation prone proteins reveals absolutely no similarity embedded within the primary sequence, which suggests that the similarity must be structural in nature. Analysis of their solved structures is futile due to the fact that the informative structure, being the amyloid state, is completely insoluble making its structural analysis by conventional high-resolution methods impossible (36). Two recent publications have attempted to shed light on this problem. First, Jonathan Blake and Fred Cohen attempted to improve alignment techniques of distant homologs with low sequence identity (37). They developed a new set of amino acid substitution matrices taking advantage of structural data to indicate evolutionary consequences of amino acid substitution. The goal of this method is to improve alignment methods for homologs of less than 30% identity. However, this technique is largely inapplicable to the question at hand as the sequence similarity between amyloid forming proteins is less than 10%, which according to the above study makes the sequence alignment indistinguishable from sheer chance. Attempts to do a pair-wise alignment with the Aβ peptide and the human Prion protein (PrP) are futile as no similarity can be found by blastp methods. Again, this points to structural similarity underlying aggregation prone proteins. The second study of particular interest by Kallberg and coworkers from the Karolinska Institute addresses the prediction issue from a structural standpoint (38). Findings that a conformational switch from a α-helix to a β-sheet triggers the aggregation event for Aβ peptide and the studied prion proteins led the investigators to pinpoint the individual structural element responsible for the switch. The authors found that while the solved structure of soluble Aβ peptide revealed that residues 15-25 comprise a helix, secondary structure prediction algorithms predicted the same residues to form a β-sheet (figure 1). The same was found to be true for the prion proteins as well as a lung surfactant protein C that has been shown to form aggregates with amyloid-like structure (46). The authors then mined the structural

database for proteins that have helical structures where β-strands are predicted. The authors found 37 proteins that fit these criteria and others detailed in the paper. The findings were validated when three of the previously unstudied proteins identified in the database search were shown to form amyloids under conditions similar to Aβ aggregation.

The technique for predicting amyloid formation described in the above paper relies on having a solved three-dimensional structure for its predictions. This is an obvious limitation as the number of sequences present in Genbank far outnumbers the structures deposited in the protein data bank. I attempted to build upon the previous findings and used the 37 proteins predicted to form amyloids as a basis for determining a possible common sequence element I chose to use the 20 proteins with the longest aberrant helix, being the helix that was predicted to be a strand, and preformed a multiple sequence alignment using clustalW with low penalties for gap creation and extension (figure 2). To my surprise, two segments of all 20 sequences aligned. Upon closer investigation the first segment corresponded to the aberrant helix in 10 of the 20 proteins. A block was constructed from block maker corresponding to this aligned segment (figure 3) in the hopes that a database search using this block would come back with the 17 proteins predicted to form amyloids not used in the alignment. However, due to the low sequence identity random proteins were matched to the block. A consensus sequence was created using emotif maker and this was used to perform a BLAST search of the swissprot database (40). The motif created was fairly general which gave thousands of matches upon a database search.

In order to refine my search, only the 10 proteins whose aligned segment corresponded to the aberrant helix were used to create an alignment and subsequent block (figure 4). Because the

block and eMotif is based largely on the three prion proteins in the alignment, the resulting database search gave primarily known homologs of the prion proteins.

Four secondary structure prediction programs were tested against each other in their ability to correctly predict a α-helix instead of the normally predicted β-sheet for three of the known amyloid forming proteins. PHD, Sspro8.0, Prof, and PsiPred all predicted β-sheets in the region of interest despite having a known structure available This is consistent with recent evaluations of the secondary structure prediction sites as current automated analyses of the techniques revealed most of the prediction algorithms to be equally good varying by only a few percentage points (39).

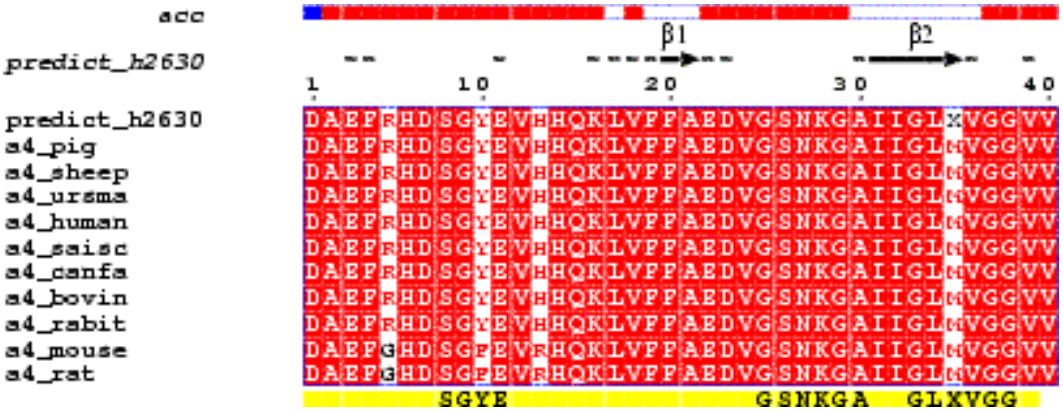
Discussion

Increases in the sensitivity of secondary structure prediction algorithms along with an increase in the numbers of sequences and structures available suggests a possibility to reveal more distant structural similarities between proteins of dissimilar function. This becomes of particular interest for proteins that have a propensity to form alternative structures off of the normal folding pathway. Amyloid forming proteins have been shown to be a main cytopathological effect in many neurodegenerative diseases. These proteins have no known sequence or functional similarity despite their common ability to form higher order aggregates. Recent reports suggest that a combinatorial approach of using low gap penalties and structure-based substitution matrices to construct multiple sequence alignments will result in identification of more distant homologs. However, due to the extreme sequence divergence between the proteins of interest, even these more sensitive techniques are not helpful for predicting amyloid forming proteins. The best approach results from a knowledge-based approach where the known biology of aggregation is used to probe the structural database looking for helices that are

predicted to be sheets. This subset of proteins reveals an underlying sequence similarity that could be useful for using a sequence-based technique for mining the database. However, even within this aligned segment of aberrant helices, no significant or useful sequence homology can be found. This reveals a dependence upon sequence identity for most of the current computational toolboxes. Emotif for example, is really only useful for alignments of sequences that are greater than 30% identical. This 30% cutoff seems to be the gold standard for most techniques giving them a bias towards merely looking for sequence homology between organisms and less useful for deciphering less obvious functional and structural similarities.

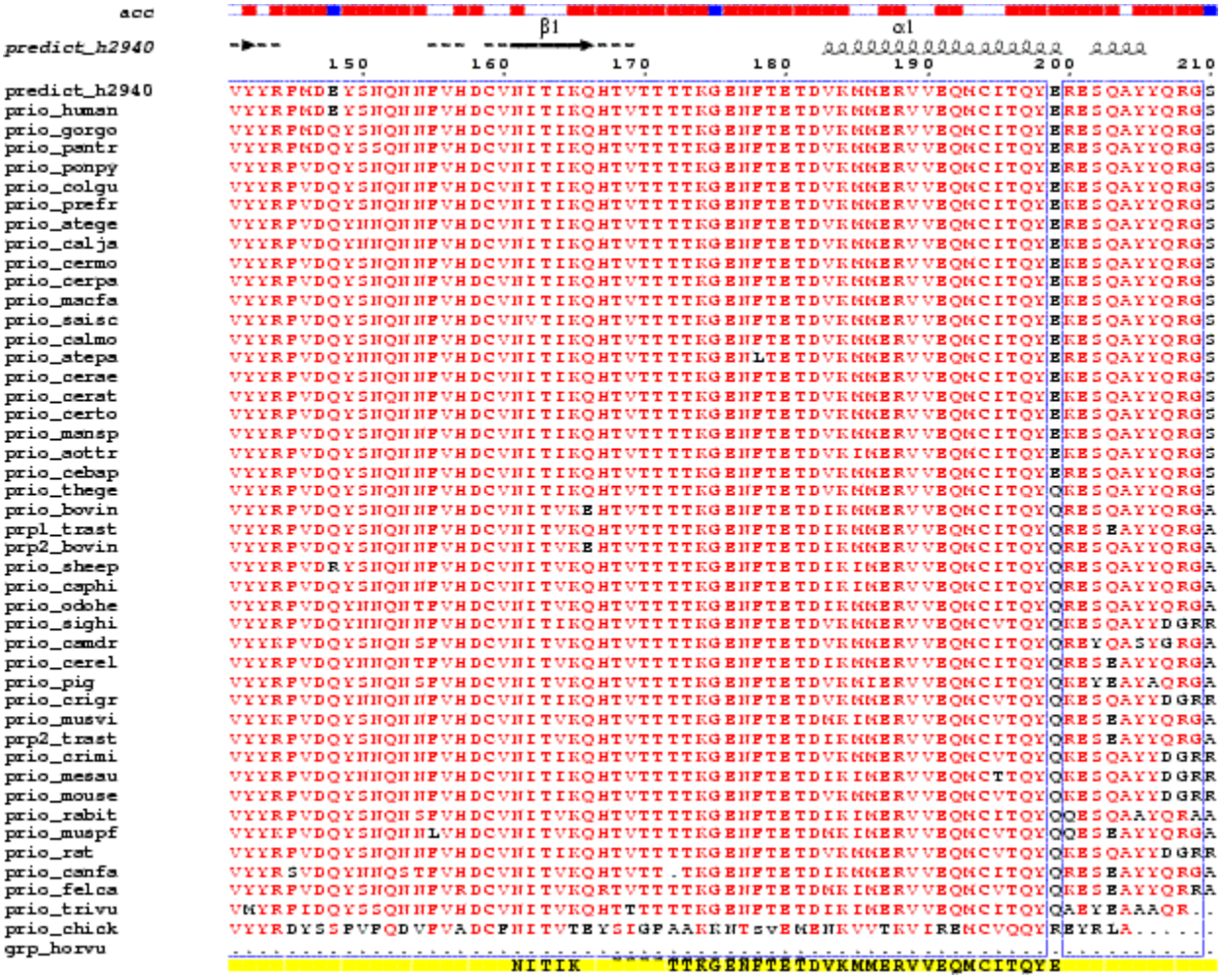
A large defect in the secondary structure prediction field is the limitation of the three state prediction placing every residue into the pedestrian categories of helix, strand, or other. A more careful categorization among prediction software and among the annotation of structures would be helpful for the classification of biological significance of structural motifs. However, with the greater complexity of classification comes an exponential increase of the computational complexity.

A



B 1 DAEFRHDSGY EVHH***QKLVFF AEDVG***SNKGA IIGLXVGGVV
SSS S S ***HHHHHT THHHH***TTTTT TTTT SS

C



D 1 GSKKRPKPGG WNTGGSRYPG QGSPGGNRYP PQGGGGWGQP HGGGWGQPHG

51 GGWGQPHGGG WGQPHGGGWG QGGGTHSQWN KPSKPKTNMK HMAGAAAAGA

```
101 VVGGLGGYML GSAMSRPIIH FGSDYEDRYY RENMHRYPNQ VYYRPMDEYS
          S EE             S HHHHHHH HHHTTTS SS  EE  SSS

151 NQ***NNFVHDCV NITIKQHTVT*** TTTKGENFTE TDVKMMERVV EQMCITQYER
     S***HHHHHHHH HHHHHHHHHH*** TGGGT    H HHHHHHHHHH HHHHHHHHHH

201 ESQAYYQRGS
    HHHHHHH
```

Figure 1: Alignment and secondary structure prediction of Aβ family and the prion family. A. Alignment and secondary structure prediction created from predictprotein (PHD) for the amyloid beta family of proteins. Predictions using other programs (Prof, PSIPRED, Sspro8.0) yielded similar predictions B. The primary sequence and annotated secondary structure from the protein data bank of the Aβ peptide. The bolded segment denotes the aberrant helix that is predicted to be a β-sheet. C. Alignment and secondary structure prediction for the prion protein family. D. The primary sequence and the annotated secondary structure for the human PrP.

```
1BCT__          ---------------MRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGAGIVPLNIETL
1SPF__          ----------------------------------------LRIPCCPVNLKRLLVVVV
2IFO__          ----------------------------------SGGGGVDVGDVVSAIQGAAGPIAA
1BA6__          ---------------------------------------DAEFRHDSGYEVHHQKLVFFA
3PTE__          IEKLTGHSVATEYQ--NRIFTPLNLTDTFYVHPDTVIPGTHANGYLTPDEAGGALVDSTE
1QLX_A          SRPIIHFGSDYEDR--YYRENMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTT
1AG2__          SRPMIHFGNDWEDR--YYRENMYRYPNQVYYRPVDQYSNQNNFVHDCVNITIKQHTVTTT
1B10_A          SRPMMHFGNDWEDR--YYRENMNRYPNQVYYRPVDQYNNQNNFVHDCVNITIKQHTVTTT
1PBV__          AFVDLHEFTDL-----NLVQALRQFLWSFRLPGEAQKIDRMMEAFAQRYCLCNPGVFQST
1MTY_D          DGFISGDAVECSLN--LQLVGEACFTNPLIVAVTEWAAANGDEITPTVFLSIETDELRHM
1VNS__          THPVVLIPVDPNNP--NGPKMPFRQYHAPFYGKTTKRFATQSEHFLADPPGLRSNADETA
1AUR_A          ASRIFLAGFSQGGA--VVFHTAFINWQGPLGGVIALSTYAPTFGDELELSASQQRIPALC
1TCA__          GLTQIVPTTNLYSA--TDEIVQPQVSNSPLDSSYLFNGKNVQAQAVCGPLFVIDHAGSLT
2OCC_K          --------------------------------IHQKRAPDFHDKYGNAVLASGATFCV
2SQC_A          LHGYQKLSVHPFRR--AAEIRALDWLLERQAGDGSWGGIQPPWFYALIALKILDMTQHPA
1WER__          KSVQHKWPTNTT----MRTRVVSGFVFLRLICPAILNPRMFNIISDSPSPIAARTLILVA
1B8O_A          QKAHSTWKQMGEQR--ELQEGTYVMLGGPNFETVAECRLLRNLGADAVGMSTVPEVIVAR
1B2V_A          AHTLYGQLDSLSFG--DGLSGGDTSPYSIQVPDVSFGGLNLSSLQAQGHDGVVHQVVYGL
1QUT_A          DEQDDPLNLKGSFA--GAMGYGQFMPSSYKQYAVDFSGDGHINLWDPVDAIGSVANYFKA
1GGT_B          VMDRAQMDLSGRGNPIKVSRVGSAMVNAKDDEGVLVGSWDNIYAYGVPPSAWTGSVDILL
```

Figure 2: Multiple sequence alignment of 20 predicted amyloid forming proteins. A multiple sequence alignment was generated using ClustalW with a Blosum62 matrix and a gap opening penalty of 10 and an extension penalty of 0.05. This represents only a subset of the total alignment with the above section representing one of only two segments where all 20 proteins aligned. This segment corresponds to the aberrant helix in 10 of the proteins.

```
ID   x13941xbl; BLOCK
AC   x13941xblA; distance from previous blocks=(431,431)
DE   ../tmp/13941.blin
BL   UNK motif;  width=13; seqs=20; 99.5%=0; strength=0
1AG2__             ( 431) NITIKQHTVTTTT  38
1B10_A             ( 431) NITIKQHTVTTTT  38
1QLX_A             ( 431) NITIKQHTVTTTT  38
1SPF__             ( 431) PVNLKRLLVVVVV  65
2IFO__             ( 431) IAAIGGAVLTVMV  60
1VNS__             ( 431) NSEVNNADFARLF  57
1BA6__             ( 431) GYEVHHQKLVFFA  81
1B8O_A             ( 431) GMSTVPEVIVARH  96
1GGT_B             ( 431) FAEVNSDLIYITA  56
2OCC_K             ( 431) TFCVAVWVYMATQ  94
1BCT__             ( 431) PLNIETLLFMVLD  57
1PBV__             ( 431) TCYVLSFAVIMLN  84
2SQC_A             ( 431) ISPVWDTGLAVLA  63
1WER__             ( 431) AKSVQNLANLVEF  82
3PTE__             ( 431) LTPDEAGGALVDS 100
1MTY_D             ( 431) FTPVLGMLFEYGS  69
1AUR_A             ( 431) CLHGQYDDVVQNA  90
1B2V_A             ( 431) HDGVVHQVVYGLM  66
1QUT_A             ( 431) HGWVKGDQVAVMA  52
1TCA__             ( 431) HAGSLTSQFSYVV  78
```

Figure 3: Block created from the above multiple sequence alignment.

The multiple sequence alignment from above was input into Blockmaker and the resulting block is represented. Two blocks were constructed from this alignment with the above block representing the segment of interest. Note the low degree of sequence similarity. Attempts to construct a consensus sequence and search the database for possible amyloid forming proteins failed due to the low sequence identity.

A

```
1BCT__          VLWSAYPVVWLIGSEGAGIVPLNIETLLFMVLDVSAKVGFGLILLRSRAIFG--------
1SPF__          LRIPCCPVNLKRLLVVVVVVVLVVVVIVGALLMGL------------------------
2IFO__          IQGAAGPIAAIGGAVLTVMVGIKVYKWVRRAM----------------------------
1BA6__          YEVHHQKLVFFAEDVGSNKGAIIGLX-VGGVV----------------------------
1QLX_A          ITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGS-----------
1AG2__          ITIKQHTVTTTTKGENFTETDVKMMERVVEQMCVTQYQKESQAYY---------------
1B10_A          ITIKQHTVTTTTKGENFTETDIKIMERVVEQMCTTQYQKESQAYYDGRRS----------
1MTY_D          VSIANDPASAKYLNTDLNNAFWTQQKYFTPVLGMLFEYGSKFKVEPWVKTWDRWVYEDWG
2OCC_K          NAVLASGATFCVAVWVYMATQIGIEWNPSPVGRVTPKEWREQ------------------
1B2V_A          GDTGALETALNGILDDYGLSVNSTFDQVAAATAVGVQHADSPELLAA------------
```

B

```
ID   x4806xbli; BLOCK
AC   x4806xbliA; distance from previous blocks=(147,147)
DE   ../tmp/4806.blin
BL   UNK motif;  width=17; seqs=10; 99.5%=0; strength=0
1AG2__             ( 147) GGYMLGSAMSRPMIHFG  34
1B10_A             ( 147) GGYMLGSAMSRPMMHFG  37
1QLX_A             ( 147) GGYMLGSAMSRPIIHFG  36
2IFO__             ( 147) GGVDVGDVVSAIQGAAG  73
1BA6__             ( 147) FRHDSGYEVHHQKLVFF  88
1MTY_D             ( 147) GQDPAGHNDARRTRTIG  78
1SPF__             ( 147) LRIPCCPVNLKRLLVVV 100
1BCT__             ( 147) NVTVVLWSAYPVVWLIG  95
1B2V_A             ( 147) GGDTSPYSIQVPDVSFG  71
2OCC_K             ( 147) NAVLASGATFCVAVWVY  96
```

Figure 4: Multiple sequence alignment and block from a subset predicted amyloid proteins. A. A multiple sequence alignment was constructed with ClustalW on the 10 predicted amyloid forming proteins whose aligned segment corresponded to the aberrant helix. B. The block that was created from the above multiple sequence alignment. Subsequent searches of the database with the created block did not retrieve any of the other 10 proteins omitted from this alignment.

.

References

1. Kelly, J. W. (1998) *Curr. Opin. Struct. Biol.* **8,** 101-106.
2. Leroux, M. R. and Hartl F.U. In *Mechanisms of protein folding. 2nd ed.* Oxford University Press, Oxford, 1999.
3. Goldberg, M. S. and Lansbury, P. T. Jr (2000) *Nature Cell Biol.* **2,** E115-E119.
4. Conway, K. A. *et al.* (2000) *Proc. Natl Acad. Sci. USA.* **97,** 571-576.
5. Sunde, M. *et al.* (1997) *J. Mol. Biol.* **273,** 729-739.
6. Serpell, L. C. (2000) *Biochimica et Biophysica Acta.* **1502,** 16-30.
7. Pan, K. M. *et al.* (1993) *Proc. Natl Acad. Sci. USA.* **90,** 10962-10966.
8. Barrow, C. J. *et al.* (1992) *J. Mol. Biol.* **225,** 1075-1093.
9. Lansbury, P. T. Jr (1999) *Proc. Natl Acad. Sci. USA.* **96,** 3342-3344.
10. Chiti, F. *et al.* (1999) *Proc. Natl Acad. Sci. USA.* **96,** 3590-3594.
11. Serpell, L. C., Blake, C. C. F., and Fraser, P. E. (2000) *Biochemistry* **39,** 13269-13275.
12. MacPhee, C. E. and Dobson, C. M. (2000) *J. Am. Chem. Soc.* **122,** 12707-12713.
13. West, M. W. *et al.* (1999) *Proc. Natl Acad. Sci. USA.* **96,** 11211-11216.
14. Chiti, F. *et al.* (2000) *EMBO J.* **19,** 1441-1449.
15. Villegas, V. *et al.* (2000) *Protein Sci.* **9,** 1700-1708.
16. Sipe, J. D. (1992) *Annu. Rev. Biochem.* **61,** 947-975.
17. Ciani, B. (2002) *J. Biol. Chem.* **277,** 10150-10155.
18. Jackson, G. S. and Clarke, A. R. (2000) *Curr. Opin. Struct. Biol.* **10,** 69-74.
19. Rochet, J. C. and Lansbury, P. T. (2000) *Curr. Opin. Struct. Biol.* **10,** 60-68.
20. Bucciantini, M. *et al.* (2002) *Nature* **416,** 507-511.
21. Perutz, M. F. *et al.* (2002) *Proc. Natl Acad. Sci. USA.* **99,** 5596-5600.
22. Chiti, F. *et al.* (2002) *Nature Struct. Biol.* **9,** 137-143.
23. Pollastri, G. *et al.* (2002) *Proteins* **47,** 228-235.
24. Rost B, and Sander C. (1994) *Proteins* **19,** 55-72.
25. Baldi, P. *et al.* (1999) *Bioinformatics* **15,** 937-946.
26. Lesk, A. M., Lo Conte, L., and Hubbard T. J. P. (2001) *Proteins* **Suppl 5,** 98-118.
27. Heyn, M. P. (1989) *Meth. Enzymol.* **172,** 575-584.
28. Przybylski, D., and Rost, B. (2002) *Proteins* **46,** 197-205.
29. Cuff, J. A., and Barton G. J. (2000) *Proteins* **40,** 502-511.
30. Jones D. T. (1999) *J. Mol. Biol.* **292,** 195-202.
31. Rost, B. and Sander, C. (1992) *Nature* **360,** 540.
32. Rost, B. and Sander, C. (1993) *J. Mol. Biol.* **232,** 584-599.
33. Pedersen, T. N. (2000) *Proteins* **41,** 17-20.
34. Barton, G. J. (1995) *Curr. Opin. Struct. Biol.* **5,** 372-376.
35. Altschul, S. *et al.* (1997) *Nuc Acids Res.* **25,** 3389-3402.
36. Mattice, W. L. (1989) *Annu. Rev. Biophys. Chem.* **18,** 93-111.
37. Blake, J. D. and Cohen, F. E. (2001) *J. Mol. Biol.* **307,** 721-735.
38. Kallberg Y. *et al.* (2001) *J. Biol. Chem.* **276,** 12945-12950.
39. Rost, B. and Eyrich, V. A. (2001) *Proteins* **Suppl 5,** 192-199.

40. Huang, J. Y., and Brutlag, D. L. (2001) *Nuc. Acids Res.* **29,** 202-204.
41. Richardson, J. S., and Richardson, D. C. (2002) *Proc. Natl Acad. Sci. USA.* **99,** 2754-2759.
42. Feldman H. J., and Hogue, C. W. V. (2002) *Proteins* **46,** 8-23.
43. Srinivasan, R and Rose, G. D. (2002) *Proteins* **47,** 489-495.
44. Gilis, D., and Rooman, M. (2001) *Proteins* **42,** 164-176.
45. Kihara, D. *et al.* (2002) *Proc. Natl Acad. Sci. USA.* **99,** 5993-5998.
46. Szyperski, T. *et al.* (1998) *Protein Sci.* **7,** 2533-2540.