SPARKS 2 and SP^3 Servers in CASP6

Hongyi Zhou[1] **and Yaoqi Zhou**[1,2]*
[1]*Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology and Biophysics, State University of New York at Buffalo, Buffalo, New York*
[2]*Department of Macromolecular Science, Key Laboratory of Molecular Engineering of Polymers, Fudan University, Shanghai, China*

ABSTRACT **Two single-method servers, SPARKS 2 and SP^3, participated in automatic-server predictions in CASP6. The overall results for all as well as detailed performance in comparative modeling targets are presented. It is shown that both SPARKS 2 and SP^3 are able to recognize their corresponding best templates for all easy comparative modeling targets. The alignment accuracy, however, is not always the best among all the servers. Possible factors are discussed. SPARKS 2 and SP^3 fold recognition servers, as well as their executables, are freely available for all academic users on http://theory.med.buffalo.edu. Proteins 2005;Suppl 7:152–156.**

Key words: fold recognition; protein threading; protein structure prediction, sequence profile

INTRODUCTION

Comparative modeling and fold recognition predict unknown structures of proteins based on solved structures of proteins. This is done by detecting close (comparative modeling) or remote (fold recognition) sequence homology between two sequences with inherent structural similarity. The template-based modeling approach will become increasingly useful for solving structures of proteins as more protein structures become available. One way to detect structural similarity is to identify close or remote sequence homology via sequence comparison. Advances have been made from the pairwise[1–7] to multiple sequence comparison,[8–12] and from sequence-to-sequence, sequence-to-profile[8,9,13] to profile-to-profile comparison.[12,14–17] Another way to detect structural similarity is via sequence-to-structure threading.[18–23] More recent work attempts to optimally combine the sequence and structure information for a more accurate/sensitive fold recognition.[7,24–36] For a recent review, see Godzik.[37] SPARKS (version 2)[36] and SP^3 (Zhou and Zhou[38]) are two fully automatic single-method servers that attempt to locate the best match between an input sequence and a known structure (single-template) from the template library of protein structures by optimally combining the sequence and structure information.

METHODS

Details about methods used in SPARKS[36] and SP^3 (Zhou and Zhou[38]) have been published elsewhere. Here, we give a brief summary for completeness. Both servers are single-method servers. They also use sequence as well as structure information for fold recognition alignment to the query sequence. The sequence information is taken from the sequence profiles of the input and template sequences that were generated from PSIBLAST.[8] The structure information is represented by a secondary structure profile (predicted versus actual secondary structures), structure-based profile-energy scoring (used only in SPARKS), and structure-derived sequence profile (used only SP^3). The total alignment score is optimized by local–local dynamic programming[39] with secondary structure–dependent gap insertion/deletion. The matches to different templates are ranked by an empirical criterion based on normalized *Z*-scores and reverse alignment scores. Only one template is used in each match. SPARKS 2 is an upgraded version of SPARKS,[36] in which the methods for parameter optimization, dynamic programming, and template ranking are from those used in SP^3.[38] The weight factors for various terms in the alignment score function and gap penalties in both SPARKS 2 and SP^3 are obtained by optimizing their respective performance in ProSup alignment benchmark.[40] The models for the top five matches ranked by SPARKS 2 or SP^3 are built by using MODELLER[41] without side-chain and loop refinement. It usually takes 30 min to a few hours to complete the fold recognition of a sequence (depending on the size of the query protein and the load of the server computer). It should be noted that both SPARKS 2 and SP^3 automatically make a weekly update of template and sequence libraries (i.e., based on new releases from the National Center for Biotechnology Information (NCBI; sequences) and the Protein Data Bank (PDB; structures), respectively.

Grant sponsor: National Institutes of Health; Grant numbers: R01 GM 966049 and R01 GM 068530. Grant sponsor: HHMI (to SUNY Buffalo). Grant sponsor: Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo. Grant sponsor: National Science Foundation of China; Grant number: 203240420391 (two-base fund to Yaoqi Zhou).

*Correspondence to: Yaoqi Zhou, Howard Hughes Medical Institute Center for Single Molecule Biophysics and Department of Physiology and Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214. E-mail: yqshou@buffalo.edu

Received 28 February 2005; Accepted 15 May 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20732

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version.

TABLE I. Results for T0269-1 by Several Top-Performing Servers

Method	Template (rank)	RMSD(Å)(Cov)[a]	GDT-score[a]	TM-score[b]	AL0_P[a]
SPARKS 2	1prxA (1)	2.15 (100%)	0.86	0.91	0.92
SP^3	1prxA (1)	2.28 (100%)	0.85	0.90	0.91
SFST	1prxB (1)	2.39 (100%)	0.85	0.89	0.89
ROBETTA	1prxA (1)	2.60 (100%)	0.86	0.88	0.90

[a] RMSD, GDT score, and correct alignment AL0_P from the official CASP6 evaluation.
[b] From the Zhang–Skolnick evaluation.[43]

TABLE II. Results for T0279-2 by Several Top-Performing Servers

Method	Template (rank)	RMSD(Å)(Cov)[a]	GDT-score[a]	TM-score[b]	AL0_P[a]
SPARKS 2	1jr2A (1)	2.64 (100%)	0.69	0.76	0.79
SP^3	1jr2A (1)	2.60 (100%)	0.69	0.76	0.74
PROTINFO	1jr2A (1)	2.72 (100%)	0.69	0.76	0.74
ACE	1jr2A (1)	2.77 (100%)	0.67	0.75	0.68

[a] RMSD, GDT score, and correct alignment AL0_P from the official CASP 6 evaluation.
[b] From the Zhang–Skolnick evaluation.[43]

RESULTS

Overall Performance

SPARKS 2 and SP^3 have been used to build the top five models for all CASP6 targets (43 comparative modeling targets and 44 fold recognition and new fold targets). The overall ranks of SP^3 and SPARKS 2 for all targets among all automatic servers are 3 and 4 (behind only metaservers ACE and ROBETTA), respectively, based on either total Global Distance Test (GDT) score of the first model given by CASP6 evaluations[42] or the total Template Modeling Score (TM-Score) of the first model given by the Zhang–Skolnick evaluation.[43] For comparative modeling (CM) targets, SPARKS 2 and SP^3 are ranked as 2 and 3 by either the total GDT score or the total TM-Score, respectively (behind Eidogen-EXPM only). The official ranking method for servers in comparative modeling is, however, based on combined *Z*-scores with penalties to those models with severe atomic clashes. Based on this ranking method, SPARKS 2 and SP^3 were ranked as the most accurate servers (1 and 2) among all servers (including metaservers) because their models have significantly fewer atomic clashes than Eidogen-EXPM. Analysis of CASP6 results indicates that SPARKS 2 and SP^3 do not always generate the best models for the comparative modeling targets. Thus, it is necessary to analyze what makes the methods successful in some targets and less so in other targets.

T0269-1 (CM/Easy)

Target T0269 is a two-domain protein, and T0269-1 is its first domain, with residues ranging from 2 to 159. Both SP^3 and SPARK 2 built the models for the whole protein. The results for Domain 1 is shown in Table I. It is clear that the top-performing servers picked either chain A or chain B of the homodimeric protein 1prx (PDB ID) whose sequence identity with T0269-1 is 27%. Thus, the key difference between different servers is in the alignment between the query sequence and the 1prx sequence. Indeed, as suggested from percent of correct alignment (AL0_P), SPARKS 2 has the best alignment, which lead to the smallest root-mean-square deviation (RMSD; 2.15 Å) from the native structure.

T0279-2 (CM/Easy)

T0279-2 is Domain 2 of the target T0279. The results of several top-performing servers are shown in Table II. Similar to T0269-1, all top-performing servers selected the same template, chain A of 1jr2 with 17% sequence identity with the target sequence. Again, the alignment accuracy is what made the modeling accuracy of SP^3 and SPARKS 2 come out slightly ahead of that given by the metaservers PROTINFO and ACE.

T0232 (CM/Hard)

The above two targets are CM/Easy targets according to the CASP6 classification. Target T0232, on the other hand, is a CM/Hard target. It is also a two-domain protein. Table III compares the results (in term of TM-Score[44]) for T0232 and its constituting domains given by several top-performing servers. SPARKS 2 gives the highest TM-Score (0.78) for T0232-1 but only a moderate TM-Score of 0.43 for T0232-2 (ranked 13) and 0.61 for T0232 as a whole protein (ranked 7 in TM-Score among all servers). Clearly, the poor performance in T0232-2 is the main reason for the moderate accuracy for the whole structure of T0232. Results from SP^3 are similar. On the other hand, the difference between the results of Domains 1 and 2 given by either LOOPP or ROBETTA is significantly smaller. It suggests that LOOPP and ROBETTA may have a better alignment score function for this target.

Another interesting observation is that different servers used different templates for T0232. This indicates that similar results can be achieved with different templates. Indeed, combinatorial extension (CE) structural alignment[45] between the native structure and template structures indicates that there are many "high-quality" templates. There are at least six templates whose Z-score $>$ 6.5 and 19 templates whose Z-score $>$ 6.0 in our template library. (Typically, proteins with a similar fold will have a

TABLE III. Results for T0232 by Several Top-Performing Servers

	Top Template		TM-score		
Method	PDB ID	CE Z-score[a]	T0232	T0232-1 (1–86)[b]	T0232-2 (91–236)[b]
SPARKS 2	1f3aA	6.6	0.61	0.78	0.43
SP^3	1glqA	6.7	0.61	0.73	0.45
LOOPP	1pgtA	6.7	0.71	0.71	0.61
ROBETTA	1aw9	6.2	0.69	0.71	0.61

[a] *Z*-score from CE structural alignment between the template and the native structure of T0232. A higher *Z*-score indicates stronger structural similarity.
[b] Residue ranges of Domains 1 and 2.

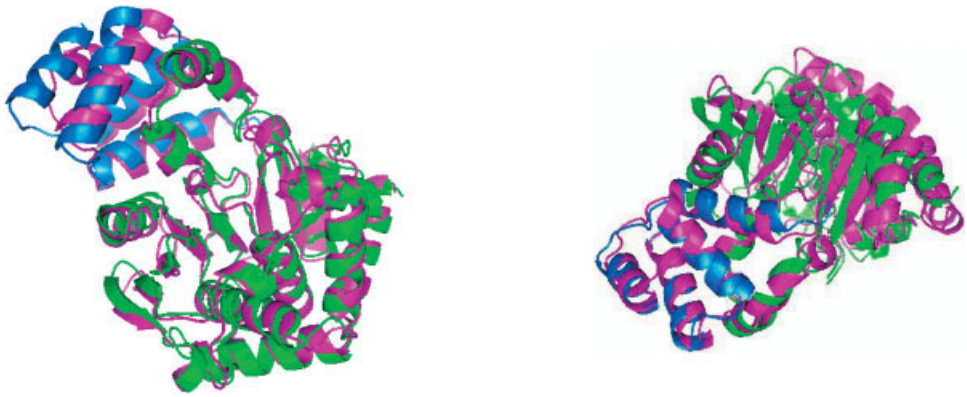

Fig. 1. Domains 1 and 2 of the native structure of T0233 are shown in blue and green, respectively. The left panel shows that Template 1v8g_A (in magenta) has a near perfect fit to Domain 2 (in green); the right panel displays the fit of Template 1o17 A (in magenta) to Domain 1 of T0233 (in blue).

Z-score of 3.5 or better.[45]) The highest structural alignment *Z*-score is 6.7. The *Z*-scores for 1f3aA (SPARKS 2), 1glqA (SP^3), 1pgtA (LOOPP), and 1aw9 (ROBETTA) are 6.6, 6.7, 6.7, and 6.2, respectively. Thus, all top-performing servers have selected "high-quality" templates. 1aw9 is in our template library but is not ranked in the top five. 1pgtA is not in our library. There are two other templates (1okta and 1duga) with a *Z*-score = 6.7 in our template library. 1duga was ranked as second in both SPARKS 2 and SP^3.

T0233 (CM/Easy)

T0233 is designated as a two-domain protein. The residue ranges for Domains 1 and 2 are 14–79 and 93–362, respectively. Domain 2 is substantially larger than Domain 1. As mentioned early, SPARKS 2 and SP^3 use only one template for one target. The template 1v8g_A (ranked 1 by both SPARKS 2 and SP^3) yielded the best server model for Domain 2, with an RMSD of 1.9Å from the native (see Fig. 1, left). However, the accuracy of this model for Domain 1 is ranked significantly lower than that from many other servers. On the other hand, template 1o17_A (ranked 3 by both SPARKS 2 and SP^3) produced the best server model for Domain 1, with an RMSD of 0.995 Å from the native (see Fig. 1, right). This example illustrates the importance of using multiple templates and/or domain prediction for accurate structure prediction.

T0235-1 (CM/Easy)

SPARKS 2 and SP^3 do not perform as well as other servers such as SFST and SAM-T99 for T0235-1, as shown in Table IV. The latter two have used the template 1nbfA, whose sequence identity with T0235-1 is 12%. The template used by SPARKS 2 and SP^3 is 1nb8A. It turns out that 1nb8A and 1nbfA share 93% sequence identity. This is why 1nbfA was not in our template library, which is based on a 40% sequence identity cutoff. However, we found that 1nbfA is a better structural template than 1nb8A, because the latter has many more missing coordinates than the former. If 1nbfA were in the template library, SPARKS 2 would yield a structure with significantly higher accuracy based on TM and GDT scores (increased by 12% and 16%, respectively). However, it is still not as accurate as those of SFST and SAM-T99. Similar results are obtained for SP^3. This indicates that there is room for further improvement of the alignment accuracy of SPARKS 2 and SP^3 for this target.

TABLE IV. Results for T0235-1 by Several Top-Performing Servers

Method	Template	GDT-score[a]	TM-score[b]
SAM-T99	1nbfA	0.69	0.81
SFST	1nbfA	0.67	0.78
SP^3	1nb8A	0.49	0.68
	(1nbfA)[c]	(0.59)[c]	(0.77)[c]
SPARKS 2	1nb8A	0.50	0.68
	(1nbfA)[c]	(0.58)[c]	(0.76)[c]

[a] GDT-score from the official CASP6 evaluation.
[b] from the Zhang–Skolnick evaluation.[43]
[c] The result if 1nbfA replaces 1nb8A as the template.

T0246 (CM/Easy)

T0246 is another CM/Easy target. SPARKS 2 and SP^3 used 1cnzA as the template (sequence identity: 55%) and achieved an RMSD of about 2.1 Å (100% coverage) from the native structure. This is slightly worse than the best server result (RMSD of 1.4Å, 98% coverage) given by 3D-JIGSAW server. The latter used 1cnzB as the template. 1cnzA and 1cnzB are two chains of a homodimer. However, their structures are not identical. The RMSD between the two structures is 1.5 Å. If 1cnzB is used as the template in SPARKS 2 or SP^3, the accuracy of predicted model is 1.6 Å, with 100% coverage. This demonstrates that even different chains of a homodimer can produce results with significantly different accuracy. We found that the total temperature B-factor of chain B (37,610) is significantly smaller than that of chain A (58,850). Thus, chain B is more rigid and its corresponding structure is likely more accurate. This explains why a more accurate model was yielded from chain B as template. It is noted, however, that the template ranking score in SPARKS 2 still ranks 1cnzA slightly above 1cnzB. This suggests that when the difference between the ranking scores of two templates is insignificant, the respective quality of the template structures themselves may be more important.

DISCUSSION

What Went Right?

The good performance of SPARKS 2 and SP^3 in CM targets is due to the consistent, high success in recognizing the best templates. For easy targets, it was nearly 100% successful in locating the best or near-best template. The only exception is T0240. SPARKS 2 (SP^3) ranked 1tol_A as second (third), although it is a better template (with a *Z*-score of 4.1 in CE structural alignment) than 1ihr_A (ranked as first by SPARKS 2 and SP^3 but with *Z*-score of only 2.3 in CE structural alignment). This exception is caused by the fact that 1ihr_A (73 residues) and T0240 (90 residues) share an identical sequence of 73 residues. Thus, the empirical template rank method based on reversed alignment, normalized *Z*-score and bonus score for structurally similar templates works well for easy targets. For harder targets, fold recognition and new fold targets, in particular, template ranking, are still challenging issues. Another reason behind the performance of SPARKS 2 and SP^3 in CASP6 is their improved alignment. This may be attributed to the optimization of parameters in SPARKS 2 and SP^3 by using an alignment benchmark.[36,38]

What Went Wrong?

The results presented here demonstrate the importance of choosing the right representative structures among homologous templates. 1nbfA is obviously better template than 1nb8A, because the latter has more missing coordinates. The fact that 1cnzB is a better template than its homodimeric chain, 1cnzA, may be associated with its significantly lower total temperature B-factor. Thus, in addition to using a cutoff of sequence identity, one should carefully select the template based on its structural quality among homologous proteins. An update for template library for SPARKS 2 and SP^3 is currently in progress.

Results from multidomain protein T0233 reveal that multiple templates and/or domain separation prior to structure prediction will likely further improve the accuracy of structure prediction. Two different templates are found to be the best match for different domains of T0233. Incorporation of domain prediction and multitemplate model building, a technique already used in some servers, such as ROBETTA and Eidogen-EXPM, will be part of the next version of SPARKS and SP^3.

ACKNOWLEDGMENTS

Our thanks to Professor Jeff Skolnick and Dr. Yang Zhang for providing us the TM-Score results, and to Dr. Chi Zhang for preparing Figure 1 for us.

REFERENCES

1. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
2. Dayhoff MO, Barker WC, Hunt LT. Establishing homologies in protein sequences. Methods Enzymol 1983;91:524–545.
3. Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. Proc Natl Acad Sci USA 1988;85:2444–2448.
4. Altschul SF, Gish W, Miller W, Myers E, Lipman D. Basic local alignment tool. J Mol Biol 1990;215:403–410.
5. Vingron M, Waterman MS. Sequence alignment and penalty choice: review of concepts, case studies and implications. J Mol Biol 1994;235:1–12.
6. Qian B, Goldstein RA. Optimization of a new score function for the generation of accurate alignments. Proteins 2002;48:605–610.
7. Teodorescu O, Galor T, Pillardy J, Elber R. Enriching the sequence substitution matrix by structural information. Proteins 2004;54:41–48.
8. Altschul SF, Madden TL, Schäffer 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;25:3389–3402.
9. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
10. Henikoff S, Henikoff JG. Amino acid substitutes matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
11. Bailey TL, Gribskov M. Score distributions for simultaneous matching to multiple motifs. J Comput Biol 1997;4:45–59.
12. Koretke KK, Russell RB, Lupas AN. Fold recognition from sequence comparisons. Proteins 2001;Suppl 5:68–75.
13. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14:755–763.
14. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–4358.
15. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles: strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
16. Yona G, Levitt M. Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. J Mol Biol 2002;315:1257–1275.
17. Marti-Renom MA, Madhusudhan M, Sali A. Alignment of protein sequences by their profiles. Protein Sci 2004;13:1071–1087.
18. Bowie JW, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
19. Godzik A, Skolnick J. Sequence–structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc Natl Acad Sci USA 1992;89:12098–12102.
20. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
21. Abagyan R, Frishman D, Argos P. Recognition of distantly related proteins through energy calculations. Proteins 1994;19:132–140.

22. Murzin AG, Bateman A. Distance homology recognition using structural classification of proteins. Proteins 1997;Suppl 1:105–112.
23. Xu Y, Xu D. Protein threading using PROSPECT: design and evaluation. Proteins 2000;40:343–354.
24. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 2001;42:319–331.
25. Yi TM, Lander ES. Recognition of related proteins by iterative template refinement (ITR). Protein Sci 1994;3:1315–1328.
26. Elofsson A, Fischer D, Rice DW, Le Grand SM, Eisenberg D. A study of combined structure/sequence profiles. Fold Des 1996;1:451–461.
27. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.
28. Rost B, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol 1997;270:471–480.
29. Jaroszewski L, Rychlewski L, Zhang B, Godzik A. Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci 1998;7:1431–1440.
30. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
31. Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.
32. Shan YB, Wang GL, Zhou HX. Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. Proteins 2001;42:23–37.
33. Al-Lazikani B, Sheinerman FB, Honig B. Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. Proc Natl Acad Sci USA 2001;98:14796–14801.
34. Kim D, Xu D, Guo J, Ellrott K, Xu Y. PROSPECT II: protein structure prediction program for the genome-scale. Protein Eng 2003;16:641–650.
35. Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B. On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J Mol Biol 2003;334:1043–1062.
36. Zhou H, Zhou Y. Single-body knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2004;55:1005–1013.
37. Godzik A. Fold recognition methods. Methods Biochem Anal 2003;44:525–546.
38. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005;58:321–328.
39. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
40. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. J Mol Biol 2000;297:1003–1013.
41. Marti-Renom M, Stuart A, 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.
42. Available online at http://predictioncenter.org/casp6/abstracts/CASP6_Tables_light.txt
43. Zhang Y, Skolnick J. Available online at http://www.bioinformatics.buffalo.edu/new_buffalo/people/zhang6/casp6/
44. Zhang Y, Skolnick J. Scoring function for the automated assessment of protein structure template quality. Proteins 2004;57:702–710.
45. Shindyalov IN, Bourne P. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.