Volume 318, number 2, 118–124 FEBS 12125 March 1993
© 1993 Federation of European Biochemical Societies 00145793/93/$6.00

The nitrogenase MoFe protein

A secondary structure prediction

Dietlind L. Gerloff[a], Thomas F. Jenny[a,b], Lukas J. Knecht[b], Gaston H. Gonnet[b] and Steven A. Benner[a]

[a]*Laboratory for Organic Chemistry* and [b]*Institute for Scientific Computation, ETH Zürich, CH-8092 Zürich, Switzerland*

Received 14 December 1992

Surface residues, interior residues, and parsing residues, together with a secondary structure derived from these, are predicted for the MoFe nitrogenase protein in advance of a crystal structure of the protein, scheduled shortly to appear in *Nature*. By publishing this prediction, we test our method for predicting the conformation of proteins from patterns in the divergent evolution of homologous protein sequences in a way that places the method 'at risk'.

Protein structure prediction; Secondary structure prediction; Nitrogenase

1. INTRODUCTION

We have recently developed procedures for extracting conformational information from patterns in the divergence and conservation in the sequences of homologous proteins [1]. These procedures are based on models for the divergent evolution of behavior and structure of proteins [2–4]. The procedures have been used to predict various aspects of the conformation of several protein families [1,5]. In the cases of protein kinase [6] and the Src homology domain 3 [7,8], secondary structure predictions were made before crystallographic data became available and shown to be remarkably accurate by subsequently determined crystal and NMR structures [9–12].

The best way to test the power of structure prediction procedures is to apply them to make predictions in advance of experimental information concerning conformation. To be useful, the predictions must be published. This ensures that knowledge of the structure cannot bias the prediction, the predictions (both correct and incorrect) are visible, and the method is placed 'at risk'. The only problem is one of coordination. A prediction published years in advance of an experimental structure is uninteresting. A prediction made even days after a structure becomes available to the predictor is useless.

In the October 29, 1992 issue of *Nature* [7], we invited scientists to send sequences to use as prediction targets for our procedure for proteins (a) the structure of which shortly will be solved, (b) where no structure is available for any obviously homologous protein, (c) where a set of homologous sequences are available, (d) where these sequences are sent to us by computer mail together with a few literature citations that provide an overview of the chemistry and biology of the protein family, and (e) when enough time is available to allow coordination of the publication of the prediction and publication of the structure. This Letter reports our first efforts directed towards this end.

Our first task has been to address challenges where criterion (e) was not fully met. For example, on November 16, Prof. D.C. Rees from the California Institute of Technology challenged us to predict a secondary structure for the MoFe protein of nitrogenase. He noted that the crystal structure of this protein had been solved, and that a manuscript coauthored with J. Kim describing that structure was in press in *Nature*, scheduled to appear in the week of December 14, 1992.

Four weeks is insufficient time to assemble a complete model for the conformation of any protein family. Nevertheless, the nitrogenase is an extremely interesting target. It is a large protein and it plays a critical role in an important metabolic process. Therefore, we have used the available time to assemble a first stage prediction of the secondary structure of this protein family. The prediction turns out to be especially instructive for those seeking to apply our procedures to their own proteins. Further, when this Letter appears in print, the issue of *Nature* containing the crystal structure will be in the library, and the success of the prediction can be immediately determined.

2. RESULTS

In presenting this prediction, we address one criticism

Correspondence address: S.A. Benner, Laboratory for Organic Chemistry, ETH Zürich, CH-8092 Zürich, Switzerland. Fax: (41) (1) 262 2437.

Published by Elsevier Science Publishers B.V.

of our procedure transmitted to us by established workers in the area: that it is inferior because it is not fully automated, and relies in part on the experimence and training of individuals making the prediction. As noted elsewhere, we do not find this criticism particularly evincing [1,2,6,12]. Conformational analysis in proteins is not fundamentally different from conformational analysis in other branches of organic chemistry, and no predictive problem in conformational analysis in chemistry has yet been solved, even for small molecules, by a fully automated procedure in the century during which conformational analysis has been developed. Rather, problems in conformational analysis are solved in chemistry by first developing a formalism. The formalism is then applied by humans to real problems. In this application, experience, training and intuition can make contributions, errors can be understood, and the formalism can be rationally improved. Organic chemical analyses can be taught, reproducibly applied, any subjected to critical testing, as any student in an undergraduate chemistry course can confirm. Of course, it is difficult to apply methods designed to evaluate automated prediction heuristics to the prediction heuristics obtained by an organic chemical paradigm. This is one reason why de novo predictions, such as the one presented here, are so important in developing the predictive formalism.

To illustrate this point, the prediction in Fig. 1 is broken into several parts. For surface, interior, parsing, and active site assignments, the first line (TJ) reports

```
             5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95  100
             .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
 a -                                                  ----------------------TINPIFTCQPAGAQFVSIGIKDCIGIVHGGQGCVMFV
 p -                                                  ----------------------IINPMYDCQPAGAQYAGIGIKDCIPLVHGGQGCTMFV
 o -                                                  ----------------------IINPMYDCQPAGAQYAGIGVKDCIPLVHGGQGCTMFV

 b -     MPQNPERTVDHVDLFKQPEYTELFENKRKNFE_GAHPPEEVERVSEWTKSWDYREKNFAREALTVNPAKGCQPVGAMFAALGFEGTLPFVQGSQGCVAYF
 c -      -QSADKVIDHFTLFRQPEYKELFERKKTEFEYGPTPTKEVARVSAWTKTEEYKEKNLPVEAVVINPTKACQPIGAMLAAQGFEGTLPFVHGSQGCVSYY
 xx-     MSQNADKIVDHFNLFKQPDYQEMFKNKQKTFE_NRLPADQVARGQEWTKTWEYREKNFAREALSVNPDKACQPLGRIFAR_GFDGTRPFVHGSQGCVAYF
 e -     MPQSAEHVLDHVELFRGPEYQQMLAKKKI_FE_NPRDPAEVERIKEWTKTAEYREKNFAREALAVNPAKACQPLGAVFASVGFERTLPFVHGSQGCVAYY
 f -     MAQSADHVLDHLELFRGPEYQQMLADKKM_FE_NPRDPAEVERIRAVTKTPEYREKNFA_EALAVNPAKACQPLGAVFVSVGFEGTLPFVHGSQGCVAYY
 d -     -------------LFLDQDYKDMLAKKRDGFEE_KYPQDKIDEVFQWTTTKEYQELNFQREALTVNPAKACQPLGAVLCALGFEKTMPYVHGSQGCVAYF
 h -     -------------LFEQDEYQELVRNKRQ_LEE_RHDAQRVQEVFAWTTTAEYEALNF_HEALTVDPAKACHALGAVLCSLGFANTLPYVHGSQGCVAYF
 i -     -------------LFEQDEYQELFRNKRQ_LEE_AHDAQRVQEVFAWTTTAEYEALNFRREALTVDPAKACQPLGAVLCSLGFANTLPYVHGSQGCVAYF

 g -                                                                RKALRINPAKTCQPVGAMYAALGIHNCLPHSHGSQGCCSYH

 k -                                                         --------KACVVNPLKMSQPIGGAYAFMGLRGAMPLLHGSQGCTSFG
 m -                                                         --------KWATINPLKSSQPLGGALAFLGVGGAIPLFHGSQGCTSFA
 n -                                                         --------RALSTNPLKTSAPLGAAMAYLGIEGAVPLFHGAQGCTAFG
 j -                                                         --------KALAVSPLKASQTMGAALAILGLALSMPLFHGSQGCTAFA
 l -                                                         --------KPLAVSPIKTGQPLGAILASLGIEHSIPLVHGAQGCSAFA
1.       .       ... .. .... ...  ..   .. . ...   ....   .. . .. . .. .......*.......*. ....*.. ......*.***...
2.TJ                                                                    sippiiiisiipiiiisipissiipiiipiipiiiii
3.SB     surface/interior position not assigned in non-conserved region IsP X i pip iIisipISSiIpII pp p i II
4.SB     surface/interior position not assigned in non-conserved region p   s s   p    P  pp     $  $$$ i
5.DG                                                                    siPPsXs/spIp XIisipISS/sPIi/pp$p$/iIi
6.TJa                                                                                                       a
7.TJb                                                                       bbbb     bbbbb                bbbbb
8.DGa                                                                                                       A
9.DGb                                                                       BBBBBbbb bbbb    BBBB         bBBBB
10.SBa
11.SBb                                                                      BBBBBb  BBBBBB   BBB bbbb     BBBBB
12.con                                                                      BBBBB   CBBBBB CC BBB      CCC

           105  110  115  120  125  130  135  140  145  150  155  160  165  170  175  180  185  190  195  200
             .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
 a -     RLIFSQHYKESFELASSSLHEDGAVFGACGRVEEAVDVLLSRYPDVKVVPIITTCSTEIIGDDVDGVIKKLNEGLLKEKFPDREVHLIA_MHTPSFVGSM
 p -     RLLFAQHFKENFDVASTSLHEESAVFGGAKRVEEGVLVLARRYPNLRVIPIITTCSTEVIGDDIEGSIRVCNRALEAE_FPDRKIYLAP_VHTPSFKGSH
 o -     RLLFAQHFKENFDVASTSLHEESAVFGGAKRVEEGVLVLARRYPELRLIPIITTCSTEVIGDDIEGTINVCNRALAAE_FPERKIYLAP_VHTPSFKGSH

 b -     RTHLSRHYKEPCSAVSSSMTEDAAVFGGLNNMIEGMQVSYQLY_KPKMIAVCTTCMAEVIGDDLGAFITNSKNAGSI_____PQDFPVPFAHTPSFVGSH
 c -     RTHLTRHFKEPNSAVSSSMTEDAAVFGGLNNMIDGLQRYALYK__PKMIAVLTTCMAEVIGDDLSGFINNAKNKESV_____PADFPVPFAHTPAFVGSH
 xx-     RSHFNRHFKEPSSCVSSSMTEDPAVFGGLNNMIDGLAISYSLY_KPKMIAVSTTCMAEVIGDDLNAFIKTAKEKGNV_____PESFDVPFAHTPSFVGSH
 e -     RSHLSRHFKEPSSCVSSSMTEDAAVFGGLNNMTDGLANSYKMY_KPKMIAVSTTCMAEVIGDDLNAFIKTSKEKGSV_____PADFDVPFAHTPAFVGSH
 f -     RSHLSRHFKEPSSCVSSSMTEDAAVFGGLNNMIDGLANSYNMY_KPKMIC_STTCMAEVIGDDLNAFIKTSKEKGSV______RRSSTPFAHTPAFVGSH
 d -     RSYFNRHFREPVSCVSDSMTEDAAVFGGQQNMKDGLQNCKATY_KPDMIAVSTTCMAEVIGDDLNAFINNSKKEGFI_____PDEFPVPFAHTPSFVGSH
 h -     GTYFNRHFKEPIACVSDSMTEDAAVFGGNNNMNLGLQNASALY_KPEIIAVSTTCMAEVIGDDLQRFIANAKKDGFV_____DSSIAWPHAHTPSFIGSH
 i -     RTYFNRHFKEPIACVSDSMTEDAAVFGGNNNMNLGLQNASALY_KPEIIAVSTTCMAEVIGDDLQAFIANAKKDGFV_____DSSIAVPHAHTPSFIGSH

 g -     RTVLSRHFKEPAMASTSSFTEGASVFGGGSNIKTAVKNIFSLY_NPDIIAVHTTCLSETLGDDLPTYISQMEDAGSI_____PEGKLVIHTNTPSYVGSH

 k -     LTLFVRHFKEAIPLQTTAMSEVATVLGGYENLEQAILNISKRA_KPKIIGICSTGVTETNGDDVE___AYLKLIRS__AYPQLTKLPLVYVSTPDFKGAF
 m -     LVLLVRHFKEAIPLQTTAMDDVAIVLGGAGHLEQAILNLKIRA_KPKLIGICTTALVETRGEDLAGDLASIKLERA___EELTGTDVVLANTPDFDGAM
 n -     VVHLVRHFKEAVPLQTTAMNEVSTILGGGEQIEEAIDNIRKRA_NPKFIGIASTALTETRGEDIAGELRAMQVRRK____DWVGTAVVHVITPDFEGGQ
 j -     KVFFVRHFREPVPLQTTAMDQVSSVMGADENVVEALKTICERQ_NPSVIGLLTTGLSETQGCDLH___TALHEFRT__QYEEYKDVPIVPVNTPDFSGCF
 l -     KVFFIQHFHDPVPLQSTAMDPTSTIMGADGNIFTALDTLCQRN_NPQAIVLLSTGLSEAQGSDIS___RVVRQFRE__EYPRHKGVAILTVNTPDFYGSM
1.       .... .*..... .  ... . ....*.. .....  ..  .. ... ... .*...*. *.*. . .   . ...   ...  . . .. ...**...*..
2.TJ     siiissiissiiiiiiiiiisiiiipiississiipiiisiippisiiipiiiiiiiiipsiisppppssisssispppppppp    ipiiipiiipii
3.SB     S iI s$Isss sIi s IsssiiIIiiSSsISSiIssISS XPSiSIIiPIi$iii$i p$$ISPPPSs sSS sPPPPPPPSSssIiiI $PpIsp
4.SB      i        ppp  i i       ppppp         si ppp  pI         s        i  s        pp pP s      ps
5.DG     S iIss$ISspppIX/s/IssXiiIIppSSsISSiISsiSS/ PPPSIIpIii$iii$i ps$ISPPPSsiSSXssPPPPPsSS SIpPIi$P ISpi
6.TJa    aaaaaaaaaaaa                AAAAAAAAAAAAAAAAA        aaaaaaaa        aaaaaaaaa                   A
7.TJb    bbbbbbb  bbbbbb     bBBBB                           bBBBBBBBB
8.DGa    AAAAAAAAAA                   aAAAAAAAAAAAAAA        aaaaaaaaaa                                  aa
9.DGb                        bBBBB                            bbbbbbb                   bbbb              b
10.SBa     AAAAAAAA                   aaaAAAAAAAAAAAAA                                                    A
11.SBb    bbbbbb             bBBBB                      bbbb bBBBBBBBb
12.con             CCC  CCC   BBBBCC  AAAAAAAAAAAA  CCC       BBBB   CCCCCCC           CCCCC      CC   CC  CCA
```

Fig. 1 (1st part)

```
          205  210  215  220  225  230  235  240  245  250  255  260  265  270  275  280  285  290  295  300
            .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
 a  - ISGYDVAVRDVVRHF___AK__REA_____________PNDKINLLTG____WVNPGDVKELKHLLGEMDIEANVLFEIES_FDSPILPD_GSAVSHGNT
 p  - VTGYAECVKSVFKTI___TDAHGKGQ___________PSGKLNVFPG____WVNPGDVVLLKRYFKEMDVEANIYMDTED_FDSPMLPN_KSIETHGRT
 o  - VTGYAECVKSMFKTI___TEVHGKGQ___________PSGKLNVFPG____WVNPGDVVLLKRYFKEMGVDATVFMDTED_FDSPMLPN_KSIETHGRT

 b  - ITGYDNMMKGILSNL___T___EGKK_____KAT___SNGKINFIPGFDTY__VGNNRELKRMMGVMGVDYTILSDSSDYFDSPNMGE_YEMYP_SGT
 c  - IVGYDNMIKGVLTHF___W_______GTSENFDTP__KNETINLIPGFDGF_AVGNNRELKRIAGLFGIQMTILSDVSDNFDTPADGE_YRMYD_GGT
 xx - ITGYDNMMKGILTHF___W__DGKAGTVPALERKP___DEKINFIGGFDGY_TVGNMREIKRLFSLMNVDYTILGDGSDVWDTPADGE_FRMYD_GGT
 e  - VTGYDNALKGILEHF___W___DGKAGTAPKLERK__PNGAINIIGGFDGY_TVGNLREIKRILELMGIQHTVLADNSEVFDTPTDGE_FRMYD_GGT
 f  - VTGYDNALKGILEHF___W___NGKAGTAPKLERK__PNEAINIIGGFDGN_TVGNLREIKRILALMGIKHTILADNSEVFDTPTDGE_FRMYD_GGT
 d  - VTGWDNMFEGIARYF___T_________LKSMDDKVVGSNKKINIVPGFETYL__GNFRVIKRMLSEMGVGYSLLSDPEEVLDTPADGQ_FRMYA_GGT
 h  - VTGWDNMFEGFAKTF___TADYQGQPGKLPKL________KLNLVPGFETYLGTGNFRVLKRMMEQMAVPCSLLSDPSEVLDTPADGQ_YRMYS_GGT
 i  - VTGWDNMFEGFAKTF___TADYQGQPGKLP__________KLNLVTGFETYL__GNFRVLKRMMEQMAVPCSLLSDPSEVLDTPADGH_YRMYS_GGT

 g  - VTGFANMVQGIVNYL___SENTGAK____________NGKINVIPGF___VGPADMREIKRLFEAMDIPYIMFPDTSGVLDGPTTGE_YKMYPEGGT

 k  - QDGWEKAVARMVEVLVD_RPSANGLRDP__________SKVNVLPGCH__LTPGDLDELRALLEDFGLYPSFLPDLAGSLDGHIPDE_FTSTTIGGI
 m  - EEGWAKAVTAMIKAITR_IGEQE__RQS__________RTIAILPGWN__LTIADIEQLRDIVESFGLKPIILPDLSGSLDGIVPDDRWVPTTYGGI
 n  - QDGWAKAVEAIVAALVP_VTAE___RDPDL________RQVTLLVPSC__FTTAEIDEAVRMIRAFGLSPIVLPDLSTSLDGHLSDD_WSGHSLGGT
 j  - ESGFAAAVKAIVETLVPERRDQVGKRP__________RQVNVLCSAN__LTPGDLEYIAESIESFGLRPLLIPDLSGSLDGHLDENRFNALTTGGL
 l  - ENGFSAVLESVIEQWVPP_APRPAQRN__________RRVNLLVSHL__CSPGDIEWLRRCVEAFGLQPIILPDLAQSMDGHLAQGDFSPLTQGGT
 1.   ..*.. ... .. .... ...  .     ..... .... .   .. .    ... . .... ..... .... . ..... .. .*..  ....  . ...
 2.TJ ispissiisiiissippppspppppppppppppppppppp    pppppppppisississiissisisisiiisisspiippiiispisiispppi
 3.SB  spISSIISSIISsI                        Sisii        PPP s SSIsSIISSIsISi IIis sSPI$ iIssSPISi SP  i
 4.SB / i                                                    ii             pps  ppp    ppp ppp    i ppp
 5.DG XspISsIISSIIS IPPPPPPPPPPPPPPPPPPPPPPPPPPPPPSIsIIPPPPPPPPPPpssSSIssiISSISIS  II s ssPI$ppsspSPsSi sPpp/
 6.TJa AAAAAAAAAAAAAAA                                        AAAAAAAAAAAA AAAAAAAAAA
 7.TJb                                                             BBBBBBBB bbbb
 8.DGa aaAAAAAAAAAAAAA                                       AAAAAAAAAAAA
 9.DGb bb                                     BBBBB                  BBBB bbb
10.SBa AAAAAAAAAAAAAAA                                       AAAAAAAAAAAA
11.SBb                                        BBBBB                  BBBBBBbbb
12.con AAAAAAAAAAAAAAA  unassigned due to gap  BBBBB CCCCCCCCC AAAAAAAAAAAA           CCCCCCCCCCCC  CCCC

          305  310  315  320  325  330  335  340  345  350  355  360  365  370  375  380  385  390  395  400
            .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
 a  - TIEDLIDTGNARATFALNRYEGTKAAEYLQKKFEIPAIIG___PTPIGIRNTDIFLQNLKKATG_KPIPQSLAH____ERGVAIDALADLTHMFLAEKRV
 p  - TVEDIADSANALATLSLARYEGNTTGELLQKTFAVPNALV___NTPYGIKNTDDMLRKIAEVTG_KEIPESLVR____ERGIALDALADLAHMFFANKKV
 o  - TVEDIADSANALATLALARYEGATTGEYLEKTFAVPNSLV___NTPYGIKNTDDMLRKIAEITG_KEIPESLVR____EPRIAWIALADLAHMFFANKKV

 b  - KLEDAADSINAKATVALQAYTTPKTREYIKTQWKQETQV____LRPFGVKGTDEFLTAVSELTG_KAIPEELEI____ERGRLVDAITD_SYAWIHGKKF
 c  - PLEATKEAVHAKATISMQEYCTPQSLQFIKREGPAGRQAY___NYPMGVTGTDELLMKLAELSG_KPSRGVKL_____ERGRLVDAIAD_SHTHLHGKRF
 xx - TFAEAEAALNAKATVCMQGISTEKTMAYIQEKGQEVVAL____HCPIGVTGTDHFLQEVSGISG_KPISEELKK____ERGRLVDAIGT_SISYLHGKKF
 e  - TLKDAANAIHAKATISMQQWCTEKTLSFAAEHGQDVLSF____NYPVGLSATDDFIVALSRISG_KEIPEQLAR____ERGRLVDAIAD_SSAHVHGKKF
 f  - HVEDTANAIHAKATISMQQWCTEKTLPFVSEHGQDVVSF____NYPVGVSATDDLLVALSRISG_KEIPEQLAR____ERGRLVDAIAD_SSAHIHGKKF
 d  - TQEEMKDAPNALNTVLLQPWHLEKTKKFVEGTWKHEVPKL___NIPMGLDWTDEFLMKVSEISG_QPIPASLTK____ERGRLVDMMTD_SHTWLHGKRF
 h  - TQQEMKEAPDAIDTLLLQPWQLLKSKKVVQEMWNQPATEV___AIPLGLAATDELLMTVSQLSG_KPIADALTL____ERGRLVDMMLD_SHTWLHGKKF
 i  - TQQEMKEAPDAIDTLLLQPWQLLKSKKVVQEMWNQPATEV___AIPLGLAATDELLMTVSQLSG_KPIADALTL____ERGRLVDMMLD_SHTWLHGKKF

 g  - KIEDLKDTGNSDLTLSLGSYASDLGAKTLEKKCKVPFKTL___RTPIGVSATDEFIMALSEATG_KEVPASIEE____ERGQLIDLMID_AQQYLQGKKV

 k  - DVDEIASMGRAGWTIAIGA_QMQRAAEVMQTKTGVPFRVF___ERLCGLHPNDDFMMFLSEISG_RPIPSKYRR____QRSQLADAMLD_AHFHIGGRKV
 m  - SVEEIRELGTAAQCIAIGE_HMRGPAEEMKTLTGVPYVLF___QSLTGLNAVDRFVSLLSSISG_RPAPAKVRR____RRAQLQDALLD_GHFHSAGKKI
 n  - RLDDIARIPRSAVTLAIGE_QMRAAAPMIEDRALVPYRVF___QSLTGLKVVDAFVRVLMELSGMQDPPPSTKR____DRARMMDAALD_AHFFTGGLRV
 j  - SVAELATAGQSVATLVVGQ_SLAGAADALAERTGVPDRRF___GMLYGLDAVDAWLMALAEISG_NPVPDRYKR____QRAQLQDAMLD_THFMLSSART
 l  - PLRQIEQMGQSLCSFAIGV_SLHRASSLLAPRCRGEVIAL___PHLMTLERCDAFIHQLAKISG_RAVPEWLER____QRGQLQDAMID_CHMWLQGQRM
 1.   ....... ... ...... .   .... ....  .. . .. ......  ..*..........* ....... .. . ..... ... *...  . ....
 2.TJ sisssssiisisiiisiispsisssisiisssi        ppppiiiisiiisiissississiippss          sssiisiiiipiiiiiiissi
 3.SB SiSSI SiiS S iIiI sP ISSisS ISSSIPPPPsSPPPPSSiIiISs $SIIsSI SI pPSSissSIsPPPPPsssXIi III$P sIIi ssSI
 4.SB      s  ppi /      is      i                 p    pppp     s  i    ppp                 i    i
 5.DG SiSSISSiISISsiIiI/pP ISS sSsISSS SsSssSppppSSpipISs/$SIISSIsSI/pPsSIpSSISSPPPPs$sXIi$III$PI/IIi//sSI
 6.TJa          aaaaaaaaa  AAAAAAAAAAAAA               AAAAAAAAAAAAAAAAA               aaaaaaaaa
 7.TJb              bbbbb                            bbbb                                   bbbb BBBBBBB  B
 8.DGa  AAAAAAAAa             AAAAAAAAAAAAAAA           AAAAAAAAAAAAAAAA                aaaaaaaaa
 9.DGb              BBBBB                                                              bbbb bbb  BBBBBB  B
10.SBa aAAAAAAAA             AAAAAAAAAAAAAAaa             aaaaAAAAAAAAAAAA               aaaaaaaaa
11.SBb              BBBBB                                                                   bbbb bbBBBbb  B
12.con  AAAAAAA  CC BBBBB  AAAAAAAAAAAAA  CCCCC   CC        AAAAAAAAAAA CCCCCC       C AAAAAAAAACCBBBBB  B
```

Fig. 1 (2nd part).

Fig. 1. Multiple alignment of the beta subfamily of the MoFe nitrogenase protein. Sequences are from the SwissProt protein sequence database using the DARWIN system. Underscores denote insertions and deletions. Dashes indicate sequences with insufficient similarity to permit alignment. Parsing strings (see text) are underlined. Proteins in subbranches in the evolutionary tree are denoted by blocks of sequences. Letters preceding lines indicate the nitrogenase with the following accession numbers in the database: a (P16267); b (P00468); c (P25314); d (P07329); e (P20621); f (P06122); g (P11347); h (P09771); i (P09772); j (P10336); k (P26507); l (P08738); m (P12781); n (P19077); o (P15334); p (P16856); xx (P15052). The highest bridge in the evolutionary tree occurs at a PAM (accepted point mutation per 100 amino acid residues) distance of 173.

Lines beginning with a number indicate the following.

Line 1: '*' for a conserved amino acid, '.' for a conserved amino acid type.

Lines 2–5: I and i designate strong and weak interior assignments. S and s designate strong and weak surface assignments. P and p designate strong and weak parsing assignments. X designates a split in polarity type. / designates a functional split. $ designates a conserved functional residue potentially part of an active site string. For discussion of these terms, see ref. 6. Line 2 shows unrefined assignments made by a computer 'expert system' on an unrefined alignment omitting sequence xx. Gaps arise from subsequent alignment refinement. Assignments are associated with a numerical probability (not indicated) that influenced the inferred secondary structures. Lines 3 and 4 show primary and secondary assignments made with computer assistance by an expert (S.A.B.) applying various heuristics by hand. Line 5 shows assignments made independently by a second expert (D.L.G.) applying various heuristics by hand.

Lines 8–11: A and a designate strong and weak α helix assignments. B and b designate strong and weak β strand assignments. α and β assignments were made independently and recorded on separate lines. TJa and TJb (lines 6 and 7) are α and β assignments made by rigorous application of secondary structure assignment heuristics using input from the expert system. DGa, DGb, SBa, and SBb (lines 8, 9, 10, and 11) are α and β assignments made by two experts (D.L.G. and S.A.B.) applying various heuristics by hand.

Line 12: a consensus secondary structure prediction to be compared with the crystal structure when it becomes available. Symbols as above, with C designating coil/turn assignments.

```
          405  410  415  420  425  430  435  440  445  450  455  460  465  470  475  480  485  490  495  500
           .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
  a -  AIYGAPDLVIGLAEFCLDLEMKPVLLLLGDDN_SKYVDD______PRIKALQENVDYG__MEIVTNADFWELENRIKNEGLELDLILGHSKGRFIS____
  p -  AIFGHPDLVLGLAQFCMEVELEPVLLLIGDDQGNKYKKD______PRIEELKNTAHFD__IEIVHNADLWELEKRINAG_LQLDLIMGHSKGRYVA____
  o -  AIFGHPDLVLGLAQFCLEVELEPVLLLIGDDQGSKYKKD______PRLQELKDAAHFD__MEIVHNADLWELEKRINDG_LQLDLIMGHSKGRYVA____

  b -  AIYGDPDLIISITSFLLEMGAEPVHILCN_NGDDTFKKEMEAI_____LAASP____FGKEAKVWIQKDLWHFRSLLFTE__PVDFFIGNSYGKYLWRDT
  c -  AVYGDPDFCLGMSKFLMELGAEPVHILST_SGSKKWEKQVQKVLDACRSASSG_________KAYGAKDLWHLRSLVFTD__KVDYIIGNSYGKYLERDT
  xx-  AIKGDPDFCLGVAGFLLELGAEPVHVVCT_RGNKDWAEKMNA______LFASSP___FGTGCHAYPGKDLWHMRSLVFTE__PVDFLIGNSYTKYLERDT
  e -  AIYGDPDLCYGLAAFLLELGAEPTHVLST_NGNKAWQEKMQAL_____LASSP____FGQGCQVYPGRDLWHMRSLLFTE__PVDFLIGNTYGKYLERDT
  f -  AIYGDPDLCYGLAAFLLELGAEPTHVLST_NGNNVAGENATL______FAGSP____FGE LPAYPGRDLWHMRSLLFTE__PVDFLIGNTHGKYLERDT
  d -  ALWGDPDFVMGLVKFLLELGCEPVHILCH_NGNKRWKKAVDAI_____LAASP____YGKNATVYIGKDLWHLRSLVFTD__KPDFMIGNSYGKFIQRDT
  h -  GLYGDPDFVMGLTRFLLELGCEPTVILSH_NANKRWQKAMNKM_____LDASP____NRARYSVFINCDLWHFRSLMFTR__QPDFMIGNSYGKFIQRVT
  i -  GLYGDPDFVMGLTRFLLELGCEPTVILSH_NANKRWQKAMNKM_____LDASP____YGRDSEVFINCDLWHFRSLMFTR__QPDFMIGNSYGKFIQRDT

  g -  ALLGDPDEIIALSKFIIELGAIPKYVVTG_TPGMKFQKEIDAMLAEAGIEGS__________KVKVEGDFFDVHQWIKNE__GVDLLISNTYGKFIAREE

  k -  AIGAEPDLLFDLSGMLHDMGAQVTVAVTTT______________QSEVIERIR________TKEVLIGDLEDLEGFAKEK__HCDLLITHSHGR______
  m -  AIAAEPDQLYQLATFFICLGAEIVAAVTTKG_____________ASKILHKVP________VEIIQVGDLGDLESLAT____HADLLVTHSHGQ______
  n -  AIGADPDLMFALSTALVSMGAEIVTAVTTTQ_____________NSALIEKMP________CAEVILGDLGDVERGAGQA__EAQILITHSHGR______
  j -  AIAADPDLLLGFDALLRSMGAHTVAAVVPAR_____________AAALVD_SP________LPSVRVGDLEDLEHAARAG__QAQLVIGNSHAL______
  l -  AIAAEGDLLAAWCDFANSQGMQPGPLVAPTG_____________HPSLRQ_LP________VERVVPGDLEDLQTLLCAH__PADLLVANSHAR______
1.     .. ...*. ... ..........  .    . .  .. .. .... .... ..... .  .*....... ........................
2.TJ   iiiispiiiiiiisiiisiiisiiiiisipppppppppppppppppppispippppppppppppiiissiiiiisiii pspppisiiiisisispppppp
3.SB   IIIIsp$ IIsIsSII SI ISiiiIIIsPSPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPisssSIX IsSXIPPSPPPIsIIIi  iisPPPPPP
4.SB       pppi        i     ppp                                       pppp   s   i  pp   pp   ppi
5.DG   IIIISp$sIIsIsSI/ SI/ISpiiIIisPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPSssssX$I/XIsSXIsPPPPS/sIII// XIsiiPPPP
6.TJa        aaaaaaaaaaaaaaaa                                                            aaaaaaaaaa
7.TJb  BBBB  bbbbbbb         BBBBBbb                                       bbbbb            BBBBbb
8.DGa            AAAAAAAAAAaaa                                           aAAAAAAAAAAA
9.DGb  BBB                   bBBBBB                                                         BBBB
10.SBa        aaaAAAAAAAAAAAA                                            aaaaAAAAAAAAa
11.SBb BBBB                   bbBBBb                                                      bbBBBB
12.con BBB CCC   AAAAAAAAAAAA BBBBB  CCCC     unassigned due to gaps         AAAAAAAA CCCC  BBBB

          505  510  515  520  525  530  535  540  545  550  555  560  565  570  575  580  585  590  595  600
           .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
  a -  ____IDYNIPMLRVGFPTYDRAGLFRYPTVGYGGAIWLAEQMANTLFADMEhKKN--------
  p -  ____IEANIPMVRVGFPTFDRAGLYRKPSIGYQGAMELGEMIANAMFAHMEYTRN--------
  o -  ____IEANIPMVRVGFPTFDRAGLYRKPSIGYQGAMELGEMIANAMFAHMEYTRN--------

  b -  _______SIPMVRIGYPLFDRHHLHRYSTLGYQGGLNILNWVVNTLLDEMDRSTNITGKTDISFDLIR
  c -  _______KIPLIRLTYPIFDRHHHHRYPTWGYQGALNVLVRILDRIFEDMDANTNIVGETDYSFDLVR
  xx-  _______GTPMIHIGFPIHDRHHHHRYPIWGYQGAVNVLVWILDRIHLELDRNTLGIGTTDFSYDLVR
  e -  _______ATPLIRIGFPIFDRHHHHRFPIWGYQGGLNVLVKILDKIFDEIDNKTNILGKTDYSFDIIR
  f -  _______GTPLIRIGFPIFDRHHHHRFPVWGYQGGLNVLVKILDKIFDEIDKKTSVLGKTDYSFDIIR
  d -  LHKGKEFEVPLIRIGFPIFDRHHLHRSTTLGYEGAMQILTTLVNSILERLDEETRGMQATDYNHDLVR
  h -  LAKGKAFEVPLIRLGFPLFDRHHLHRQTTWGYEGAMNIVTTLVNAVLEKLDSDTSQLAKTDYSFDLVR
  i -  LAKGKAFEVPLIRLGFPLFDRHHLHRQTTWGYEGAMNIVTTLVNAVLEKLDSDTSQLGKTDYSFDLVR

  g -  _______NIPFVRFGFPIMDRYGHYYNPKVGYKGAIRLVEEITNVILDKIERE----------

  k -  _QAAGRLKVPFYRVGFPIFDRLGAGHQVSVGYRGTRNVIFQIANLVIAHRD-----------------
  m -  _HASARLGTPLMRVGFPVFDQLGSQHKLTILYHGTRDLIFEVSNIFQSH-------------------
  n -  _HAAAALHLPLVRAGFPIFDRIGAQDTCRIGYRGTRAFFFEIANAMQA--------------------
  j -  _ASARRLGVPLLRAGFPQYDLLGGFQRCWSGYRGSSQVLFDLANLLVEH-------------------
  l -  _DLAEQFALPLVRAGFPLFDKLGEFRRVRQGYSGMRDTLFELANLIRER-------------------
1.     .. .  ...*..*...*..*. .... .. .* *.  .. .... .. ...     .  ......... .
2.TJ   pppppppsipiiiiiipiiisiiiissisiiispiisiiisiipsiiipppppppppppppppppppp
3.SB   PPPPPPPSIpII IiIpII$si     Si iiISpi SIIXSIIsXI SPPPPPPPPPPPPPPPPPPP
4.SB                            is pp    ipi   s    s i
5.DG   PPPPPPPSiPII IpIPII$sI/i sS siIISpiXSIIsSIissI SSPPPPPPPPPPPPPPPPPP
6.TJa                                    AAAAAAAAAAAAAAA
7.TJb            BBBBBB      bbbb   bbbb
8.DGa            aaaaaaaaaaaaa         aaAAAAAAAAAAAAAAA
9.DGb            BBBBBB     BBBB  BBBBbb
10.SBa                                   AAAAAAAAAAAAAAA
11.SBb           BBBBBB   BBBBB     BBBB
12.con unassignedBBBBBB     BBB     BBBB AAAAAAAAAAAAAAA    unassigned
```

Fig. 1 (3rd part).

those made by a fully automated package that is essentially an 'expert system' attempting to reproduce assignments made by organic chemists using experience, training and intuition applying procedures described in detail elsewhere [1,2,5,6,11]. This is the first time this package has been applied. The second, third and fourth lines reflect two sets of predictions prepared independently by two experts (D.G. and S.B.). A comparison of these lines illustrates the range of assignments made when relying on the experience, training, and intuition of individual scientists.

Secondary structure predictions, derived from patterns in surface and interior assignments, are likewise assigned separately, first by a rigorously applied heuristic (TJa and TJb, for α and β assignments) and then by two experts acting independently. The final line contains the consensus of all three predictions resulting from discussion among the experts, with the computer prediction represented by an expert (T.J.) as well. Special emphasis was placed on identifying core secondary structural units, as these are the most critical in assembling a tertiary structure model. Finally, an additional sequence (labeled xx) was introduced later into the multiple alignment to illustrate the extent to which assignments might be altered by additional sequence information.

A new procedure was used to help identify 'breaks' (or 'parses') in the secondary structure of a protein. In this procedure, dipeptides in the sequence composed of Pro, Gly, Asp, Asn, Ser, or any combinations of these were identified as 'parsing strings'. Further description of the use of parsing strings as indicators of breaks in secondary structure will be presented elsewhere.

3. DISCUSSION

The first stage prediction used a multiple alignment of one family (the β family) of the MoFe protein of nitrogenases only. A second stage prediction would include input from the second, more distantly homologous, α family, which aligns satisfactorily over part of the sequence. Preliminary study of the α family yielded secondary structure predictions that strongly confirm several predictions made in the first family (e.g., the α helix assigned to positions 131–142). The comparison does not, however, help define the conformation of the unusually structured (yet certainly important, judging by a variety of sequence features) stretch from positions 165–200.

Further, the alignment was subjected only to minimal revision. In a second stage prediction, revised versions of the multiple alignment would be considered in an effort to optimize secondary structural assignments. Further, in this first stage prediction, neither a supersecondary nor a tertiary structure was modeled, nor did we use information available regarding the active site of the enzyme, the subunit structure, or the biological function of this enzyme [13]. These procedures often help identify errors in the secondary structure prediction [6]. There was, regrettably, too little time.

A certain number of inconsistencies can undoubtedly be found in the figure, again due to a shortage of time. The authors welcome inquiries, as well as additional sequences for prediction.

NOTE ADDED IN PROOF: *JANUARY 4, 1992*

At the Editor's request, we have compiled recently published crystallographic data for the MoFe nitrogenase protein from *Azotobacter vinelandii* [14] in a form that allows them to be compared with a first stage prediction for the protein family (Fig. 1), completed before the crystallographic data were available. Three points are important.

First, we normally do not publish discussions of our own predictions [12] until after they have been evaluated by others. Premature evaluations by predictors of their own predictions encourage a certain type of criticism that can obscure important science, no matter how circumspect these evaluations might be. Thus, our prediction of protein kinase [6] was evaluated first by the crystallographers who solved the structure [9], by Thornton et al. [15], and then briefly by Lesk and Boswell [16]. For the SH3 domain prediction, a summary of the prediction was evaluated by Sander [17] (the prediction paper was not available to the evaluators when they made their evaluation); an editorial evaluation of the full prediction will appear simultaneously with the prediction paper [8].

Second, our central message [1] is that the organic chemist's research strategy, where a scientist actively applies a chemical formalism during the prediction process, is more likely to yield useful results than one focusing on obtaining automated computational methods. This means that methods designed to evaluated automated predictions are often deceptive when applied to predictions made using other research paradigms. With a prediction method based on a chemical formalism, it is appropriate to ask *why* a secondary structure assignment is correct (if it is correct), or why iti s incorrect (if it is incorrect). This is especially true for a first stage prediction (Fig. 1). Fig. 2 shows several points where the prediction was influenced by gaps, problematic alignments, ambiguous patterns in surface and interior assignments, and other issues often resolved during refinement (reference [6] discusses refinement procedures). As noted above, there was insufficient time to address any of these issues.

Third, evaluating predictions made from multiple alignments raises issues that are central to the field, not peripheral as this short note might imply. A structural model for a family of proteins does not apply exactly to any individual family member, and it is not always clear how to correlate a 'consensus' model to the conformation of an individual protein. It is clear, however, that consensus models are best evaluated using more than one experimental structure, as illustrated by the example of the SH3 domain [10,11].

Overall, the results for the MoFe nitrogenase protein are typical for a first stage unrefined prediction. Helix assignments are normally rather accurate; β-strands are less so. Problems are often encountered in unrefined predictions when assigning secondary structure near the active site (e.g. the first line of Fig. 2). Here sequence divergence is dominated by functional constraints relating to catalytic function, obscuring patterns that indicate particular types of secondary structure.

We ourselves evaluate a first stage prediction by grouping the assigned units in 7 categories: 'correct' (a predicted secondary structure unit that would not adversely affect an effort to build a tertiary structure model), 'possibly correct' (a predicted secondary structure unit whose effect on a tertiary structure model depends on context), 'wrong' (a helix assigned as a strand, tabulated as an incorrect strand assignment, or a strand assigned as a helix, an incorrect helix assignment), 'missed significant' (a helix or strand not identified in a region that does not contain a gap, and where the missed unit is important to a tertiary structural model), 'missed insignificant' (a helix or strand not identified in a region that does not contain a gap, but where the missed unit does not appear important to building a tertiary structure), 'gapped' (a helix or strand

```
Align #   65   70   75   80   85   90   95   100  105  110  115  120  125  130
           .    |    .    |    .    |    .    |    .    |    .    |    .    |
Seq       TVNPAKACQPLGAVLCALGFEKTMPYVHGSQGCVAYFRSYFNRHFREPVSCVSDSMTEDAAVFGGQQ
Predict         BBBBB  CBBBBB CC BBB     CCC              CCC  CCC    BBBBCC  A
Cryst   ..BBB     AAAAAAAAAAA   BBBBBBB AAAAAAAAAAAAAA    BBBBBBB  AAAAAA  AA
            .    |    .    |    .    |    .    |    .    |    .    |    .
Cryst #         70        80        90        100       110       120

Align #         140  145  150  155  160  165  170  175  180  185  190  195  200
            .    |    .    |    .    |    .    |    .    |    .    |    .    |
Seq     NMKDGLQNCKATY_KPDMIAVSTTCMAEVIGDDLNAFINNSKKEGFI_____PDEFPVPFAHTPSFVGSH
Predict AAAAAAAAAAA   CC       BBBB   CCCCCCC         CCCCC      CC   CC  CCA
Cryst   AAAAAAAAAAAAA   BBBBBBBBAAAAAA   AAAAAAAAAAA           BBBBBBBBB     A
        |    .    |     .    |    .    |    .    |    .         |    .    |
Cryst #130       140        150       160       170            180       190

Align #    205  210  215  220  225  230  235  240  245  250  255  260  265  270
            .    |    .    |    .    |    .    |    .    |    .    |    .    |
Seq     VTGWDNMFEGIARYF____T_________LKSMDDKVVGSNKKINIVPGFETYL__GNFRVIKRMLSEMG
Predict AAAAAAAAAAAAAAA  unassigned due to gaps   BBBBB CCCCCCCCC AAAAAAAAAAAA
Cryst   AAAAAAAAAAAAAAA    A                      BBBBBBB    A  AAAAAAAAAAAAA
         .    |    .                 |    .    |    .    |      .    |    .
Cryst #      200                    210       220       230         240

Align #    275  280  285  290  295  300  305  310  315  320  325  330  335  340
            .    |    .    |    .    |    .    |    .    |    .    |    .    |
Seq     VGYSLLSDPEEVLDTPADGQ_FRMYA_GGTTQEEMKDAPNALNTVLLQPWHLEKTKKFVEGTWKHEVPKL
Predict            CCCCCCCCCCCC  CCCC  AAAAAAA  CC BBBBB  AAAAAAAAAAAAA  CCCCC
Cryst    BBBBBBB AAAA                  AAAAAAAAA  BBBBB     AAAAAAAAAAA BBBBBB
         |    .    |     .    |     .    |    .    |    .    |    .    |    .
Cryst # 250       260        270        280       290       300       310

Align #    345  350  355  360  365  370  375  380  385  390  395  400  405  410
            .    |    .    |    .    |    .    |    .    |    .    |    .    |
Seq     ___NIPMGLDWTDEFLMKVSEISG_QPIPASLTK____ERGRLVDMMTD_SHTWLHGKRFALWGDPDFVM
Predict    CC        AAAAAAAAAAA CCCCCC       C AAAAAAAAACCBBBBB   BBBB CCC
Cryst      B     AAAAAAAAAAAAAA      AAAAA    AAAAAAAAAAA AAAAA  BBBBBBB AAAAA
              |    .    |    .     |    .        |    .     |    .    |    .
Cryst #      320       330        340           350        360       370

Align #    415  420  425  430  435  440  445  450  455  460  465  470  475  480
            .    |    .    |    .    |    .    |    .    |    .    |    .    |
Seq     GLVKFLLELGCEPVHILCH_NGNKRWKKAVDAI_____LAASP____YGKNATVYIGKDLWHLRSLVFTD
Predic  AAAAAAAAAAAA BBBBB  CCCC  unassigned    due    to    gaps   AAAAAAAA C
Cryst   AAAAAAAAA BBBBBBBB     AAAAAAAAAA     AAA         BBBBBB    AAAAAAAAA
          |    .    |    .     |    .         |        .    |    .    |    .
Cryst #  380       390        400            410           420       430

Align #    485  490  495  500  505  510  515  520  525  530  535  540  545
            .    |    .    |    .    |    .    |    .    |    .    |    .
Seq     __KPDFMIGNSYGKFIQRDTLHKGKEFEVPLIRIGFPIFDRHHLHRSTTLGYEGAMQILTTLVNSILE
Predic  CCC  BBBB      unass. gaps    BBBBBB    BBB     BBBB AAAAAAAAAAAAAAA
Cryst       BBBBBBAAAAAAAAAAAAAA   BBBBBBB        AAAA      AAAAAAAAAAAAAAAA
            |    .    |    .    |    .    |    .    |    .    |    .   |
Cryst #    440       450       460       470       480       490      500
```

Fig. 2. The sequence of the MoFe nitrogenase protein from *Azotobacter vinelandii*, numbered according to the multiple alignment in Fig. 1, followed by the first stage, unrefined secondary structure prediction (Fig. 1) and the secondary structure assigned by crystallography [14]. A = α helix, B = β strand, C = coil or turn. Beneath is the sequence numbering of the MoFe nitrogenase protein from *Azotobacter vinelandii*, the protein the crystal structure of which was solved (sequence *d* in the multiple alignment in Fig. 1). Positions not designated A, B, or C in the prediction are left blank; non-assignments are 'canonical' in a first stage prediction whenever the multiple alignment includes a gap and whenever the 'expert' assignments disagree. See references [1], [6] and [12] for further discussion of canonical assignments in a first stage prediction and procedures used for refining these predictions.

Table I

Secondary structure of the MoFe nitrogenase protein: comparison of the prediction and the crystal structure

	α helices	β strands
Correct	10	7
Possibly correct	0	2
Wrong	0	3
Missed significant	3	4
Missed insignificant	3	0
Gapped	2	1
Overpredicted	0	2

not identified because of the canonical treatment of gaps [6,12]), and 'overpredicted' (a helix or strand assigned to a region left unassigned by the experimentalists). These numbers for the MoFe nitrogenase protein are collected in Table I. Note that these are preliminary assignments; precise assignments can be made only in the context of an effort to assemble a tertiary structure model, which necessarily follows refinement.

Above all, this comparison illustrates the importance of early communication between crystallographer and predictor to ensure that adequate time is available for refinement. We are unable to say how much our prediction would have been improved by refinement. However, adjustments made to the multiple alignment, a standard part of a refinement procedure, should at least have allowed detection of some of the secondary structures in the regions left unassigned due to gaps (see Fig. 2). More challenging would have been improvement of the secondary structure prediction in the region of the active site.

REFERENCES

[1] Benner, S.A. (1989) Adv. Enzyme Regul. 28, 219–236.

[2] Benner, S.A. and Ellington, A.D. (1990) Bioorg. Chem. Front. 1, 1–70.

[3] Gonnet, G.H. and Benner, S.A. (1991) Computational Biochemistry Research at ETH, Technical Report 154, Departement Informatik, ETH Zürich, Switzerland.

[4] Gonnet, G.H., Cohen, M.A. and Benner, S.A. (1992) Science 256, 1443–1445.

[5] Benner, S.A., Cohen, M.A., Gonnet, G.H., Berkowitz, D.B. and Johnsson, K., in: The RNA World (R. Gesteland and J. Atkins, Eds.), Cold Spring Harbor Press, in press.

[6] Benner, S.A. and Gerloff, D. (1991) Adv. Enzyme Regul. 31, 121–181.

[7] Benner, S.A., Cohen, M.A. and Gerloff, D. (1992) Nature 359, 781.

[8] Benner, S.A., Cohen, M.A. and Gerloff, D. (1992) J. Mol. Biol , in press.

[9] Knighton, D.R., Zheng, J., Ten Eyck, L.F., Ashford, V.A., Xuong, N.H., Taylor, S.S. and Sowadski, J.M. (1991) Science 253, 407–414.

[10] Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. and Saraste, M. (1992) Nature 359, 851–855.

[11] Yu, H., Rosen, M.K., Shin, T.B., Seidel-Dugan, C., Brugge, J.S. and Schreiber, S.L. (1992) Science 258, 1665–1668.

[12] Benner, S.A. (1992) Curr. Opin. Struct. Biol. 2, 402–412.

[13] Kim, J. and Rees, D.C. (1992) Science 257, 1677–1682.

[14] Kim, J. and Rees, D.C. (1992) Nature 360, 553–560.

[15] Thornton, J.M., Flores, T.P., Jones, D.T. and Swindells, M.B. (1991) Nature 354, 105–106.

[16] Lesk, A.M. and Boswell. D.R. (1992) BioEssays 14, 407–410.

[17] Rost, B. and Sander, C. (1992) Nature 360, 540.