PATTERNS OF DIVERGENCE IN HOMOLOGOUS PROTEINS AS INDICATORS OF SECONDARY AND TERTIARY STRUCTURE: A PREDICTION OF THE STRUCTURE OF THE CATALYTIC DOMAIN OF PROTEIN KINASES

STEVEN A. BENNER and DIETLINDE GERLOFF
Laboratory for Organic Chemistry, E.T.H., CH-8092 Zurich, Switzerland

INTRODUCTION

The protein kinases form a large family of homologous proteins, with the sequences of nearly 100 protein kinase catalytic domains available from work in many laboratories (1). This family of proteins has considerable medical interest (2, 3), as many of its members are encoded by oncogenes and are therefore presumed to be intimately involved in the development of cancers. Further, protein kinases are believed to provide the physiologically relevant binding sites for tumor promoters. Finally, kinases in healthy organisms stand astride regulatory pathways important in metabolism at all levels. Much remains to be learned about this fascinating family of proteins. In particular, the three dimensional structure is not yet known for the catalytic domain in any member of the family, information that is a prerequisite for any analysis of the behavior of these proteins at a molecular level.

Recently on these pages we outlined an approach for predicting the folded structure of enzymes from the sequences of a set of aligned homologous proteins (4, 5). This approach emerged as a logical consequence of two chemical and evolutionary assumptions: (a) conformational *instability* is an evolutionarily selected trait in natural proteins, and (b) the number of conformationally stabilizing interactions available to a normal enzyme is far greater than necessary to achieve the desired level of conformational stability in the environment for which most enzymes are adapted.

These assumptions suggest that even if rules (e.g., a 'protein folding code') (6) were to exist that join constitution and conformation in proteins, globular proteins would have evolved to violate them to achieve the desired level of instability, making it essentially impossible to predict a tertiary structure from a single sequence using one of the 'metalanguages' that has emerged in the field over the past two decades to abstract the details of a folded polypeptide chain (7). While it is, of course, axiomatic in chemistry that the conformation of any molecule can be predicted by a process that explicitly evaluates the conformational energy of all conformational states, it seems unlikely that computational power will increase sufficiently fast to make direct computation of conformational energies of enzymes possible in the near future.

This logic compelled us to develop an approach different from those currently in use for modeling secondary and tertiary structure in proteins (for a review of classical methods, see Fasman, Reference 7). Our approach extracts information regarding the secondary, tertiary, and quaternary structure of a protein from the pattern of conservation and divergence in a set of its homologs, and is based on the well known fact that tertiary structure diverges far more slowly than primary structure in homologous proteins (8).

Our approach has the following features:

(a) Algorithms are designed to reflect the fact that sequence divergence within a set of homologous proteins reflects two contrasting evolutionary processes, natural selection and neutral drift.

(b) These algorithms assign positions in the alignment of homologous proteins to the surface of the protein, to the interior, to parsing segments, and to the active site.

(c) Different versions of these algorithms generate sets of structural assignments with varying degrees of reliability, beginning with versions that make a small number of highly reliable assignments and proceeding to those that make larger numbers of assignments with lower reliability.

(d) Computer tools are used to assist the biological chemist in developing an intuition for, and an organic chemical understanding of, the family of proteins being examined. Secondary and tertiary structural modeling is, however, done by the chemist himself, based on this understanding and intuition, an approach similar to that used by a chemist to analyze the conformation of much smaller organic molecules.

(e) Biological and biochemical information unique to the set of proteins being considered is used to assist the modeling. The best structural modeling is done by biological chemists who understand the biochemistry of the system they are studying, and use what they know in the modeling effort.

Our approach has some precedence in the literature, in particular recent work by Patthy (9), parts of the pattern matching approach being explored by Cohen and others (10), and the very recent work of Thornton, Sternberg, Blundell, and their co-workers. Nevertheless, several aspects of the approach remain controversial. In particular, it is the goal of many biochemists to obtain a distributable computer program that will automatically generate a secondary (or, ideally, a tertiary) model of the protein, and it is often suggested that this is the only acceptable

goal of research in this area. We believe that this goal is unrealistic given the present state of knowledge and plausible assumptions about how this knowledge is likely to develop in the near future. Further, we have been especially deterred from pursuing this goal by the knowledge that it has not yet been attained for the conformational analysis of much smaller molecules in aqueous solution.

Our approach has been tested on alignments of proteins with known crystal structures, where its performance has been satisfactory. The surface and interior algorithms, for example, make assignments that are well over 90% accurate, with the accuracy increasing with increasing number of aligned sequences (4, 5).

These assignments often allow an experienced biological chemist to identify regions of secondary structure within a protein; the identification of surface helices and internal beta strands is most successful. Secondary structural predictions based on the assignments typically identify better than 90% of the secondary structural units in a protein, with the residue-by-residue assignments of secondary structure correct over 70% of the time. The principal shortcomings of our approach are occasional misassignments of surface beta strands as surface coils (and vice versa), difficulties in assigning a conformation to strings near the active site (where functional conservation and adaptive variation often obscure patterns of conservation and variation that might otherwise indicate a particular secondary structure), and difficulties defining precisely the beginnings and ends of the secondary structural elements. As homologous proteins have *similar*, but not *identical* structures, it is likely that these results are close to being the best that can be obtained using a method based on a comparison of homologous sequences.

If the secondary structural units form a recognizable supersecondary pattern, or fall into a particular taxonomical class of protein fold (11), or if a small number (3–5) of distance constraints are available (for example, from disulfide bonds or chemical modification studies), a medium resolution tertiary structure can often be proposed for the enzymes in question. For example, a medium resolution structure of ribonuclease A (RNase) can be built from an alignment of fewer than 40 proteins using this approach.

It is difficult, of course, to evaluate the generality of our approach by applying it to the 'prediction' of the conformation of a family of proteins when a crystallographic analysis is already available for one member of the family. Knowledge of a structure will almost certainly influence a human biochemist attempting to build a model of a structure. There are three ways in which this difficulty might be overcome. First, one might implement a fully computational 'expert system' which reproduces the understanding and intuition of the biological chemist. This is, of course, challenging. Second, one may teach the approach to students to learn whether they can, in the absence of knowledge about a specific protein's structure, reproduce the structural prediction made by the biological chemist. This attempt has been under way at the E.T.H. for several years with some success. Finally, one might attempt a true prediction, to build secondary and tertiary structural models for proteins where no member of the homologous family has been studied crystallographically. This is the approach that is the focus of this paper.

To test our approach requires a protein family that meets three criteria: (a) there are at least 10 alignable sequences that are satisfactorily distributed over an evolutionary tree; (b) no crystal structure is presently available for any member of the family; and (c) a crystal structure of a member of the family is likely to become available in the not-too-distant future.

The first criterion is more than adequately met by the protein kinase family introduced in the first paragraph of this paper. An alignment of 79 sequences of homologous protein kinases (Fig. 1) displays the full range of sequence divergence required by our approach. This is more than twice the

```
protein # ->  5    10     15     20    25    30     35    40     45     50    55     60     65    70     75    80     85    90
              |    |      |      |     |     |      |     |      |      |     |      |      |     |      |     |      |     |
position
001   29   DDGEKQHSS TTT-STST EDEEFDA GGGG AEESM DNQDTMGS SSSSEEG EESEEEEESI TTSEEE TTA EEEEDSRL DNNK
002   30   QQDDNDDDD DDDDDDDD EENNSVK NVDD NNDRE DEDERGSQ EEEEQQQ SSSSSTTSEQ DDEDDD EEE KKNKKQDI QNRN
003   31   FFFFFFFFF FFFFFFFF YYYYMYY YWWW YYYYY YYLLFFLY VVVVVVV LLILILLLIL IIILLV FLL IIILLLIV LLLL
004   32   EDEEQQQQN NNNNSNSN QQEENND QREE KQTET QKVERTRQ MQQQCCC RRTQKKKKQM TMIVVL KRR TTITNKVH VQSV
005   33   RRRRIIIII FFFFFFFF LLPPSII ILFF RKKPK LLQFNIFV LLLLLLL LLLLLLLLLL MMMLLL KKK LMQLLLLF IFFL
006   34   IIKKLLMLI LLLLLLLI FYKKKEK VGVL LVIVE VILLVHVG SLLLLML EEEIVVVELM KKHGGG IVG LSLRHLKN GGGG
007   35   KKKKRRRRD MMMMMMMK EEEEEEA KKEE EEEAK RDGDTGSP TKKKQHH VVRKKEEKRE HHNEEE KKG RRALKRWE RKKK
008   36   TTTTTTTTT VVVVVVVV EDIIARL TTTT KKKEK KKKELAIL RRRRRRR KKRRRRRKKE KKKQQR VVV EEPLLFEV TTTT
009   37   LLLLLLLLL LLLLLLLL LILLLLI LLVV VIIIV LIILLLIL IIIILLL LLLLLLLLLL LLLIII LLL LLLLLLLI LLLL
010   38   GGGGGGGGG GGGGGGGG GGGGGGG GGGG GGGGG GGGGGTGG GGGGGGG GGGGGGGGGG GGGGGG GGG GGGGGGGG GGGG
011   39   TTTTTTTTV KKKKKKKK KKRRGSR ETAA EEEVE REAHSPAS STTTASS QQTNAAAARS GGGRRR SSM QQQSSSER SAAE
012   40   GGGGGGGGG GGGGGGGG GGGGGGG GGGG GGGGG GGGGGGGG GGGGGGG GGGGGGGGGG GGGGGG GGG GGGGGGGG GGGG
013   41   SSSSSSSSG SSSSSSSS AAVVKKS SSSS TTTAT KTNNESAG SSSSGGG CCCQQQQQNQ QQQNNN AAA SSSAAAAH AAAE
014   42   FFFFFFFFF FFFFFFFF FFSSFFF FTMM YYYYY YFSYFEYF FFFFFFF FFFFFFFFFF YYYFFF FFF FFFFFFFF FFFF
015   43   GGGGGGGGG GGGGGGGG SSSSGGS GSGG GGGGA SSGGSGGG GGGGGGG GGGGGGGGGG GGGGGG GGG GGGGGGGG GGGG
016   44   RRRRRRRRR KKKKKKKK VVVVAQR KCKK VVVTV ESTNECVS TTTTSSS EEDEEREEEV EEDEEE TTR MMMEEEKC QKKK
017   45   VVVVVVVVV VVVVVVVV VVVVVVV VVVV VVVVV VVVVVVVV VVVVVVV VVVVVCVVVV VVVVVV VVV VVVVVVVV VVVV
018   46   MMMMHHHHE MMMMMMML RRRRCFV KRKK YYYYY FYVSFFYY YFFFYYY WWWWWGWWFR YYYFFF YYY YYYYYYFY VVVV
019   47   LLLLLLLLL LLLLLLLL RRRRTRR LLLL KKKKL EKKKQDKS KRRRKKK MMLMMWMMYR EEESSS KKK EEEEEELH EEEK
020   48   VVVVIIVIV ASSSAAAA CCCCCLV AAAV AAGAG AAAVVSAG GGGGAAA GGGGGGGAGG GAGGGG GGG GGGGGGAG AAAA
021   49   KKKKRRRRQ DEEEEDEE VVIITVE YKKK LRRRC IKLLESEI KRRLTTT TTTTYTYTKK VVYRRR LIV NVITTQET TTTT
022   50   HHHHSSSSL RRRRRRRR KKHHEEH HHHH DHHDQ NDHHDHDR WWWWYYY WWWWYTYYWW WWWLLL WWW AALAALCL AAAA
023   51   MKKKRRVNK KKKKRKRK VLKKKKR TARR LKKPH IIVKPPIV HHHHRHH NNNNNTNNRR KKKRRR IIV RKKVLKHL HFYF
024   52   EEAAHHHHS GGGGGGGG LCPPSKA TKYQ RLTHS TTPPVDYA GGGGGGG GGGGNGGKNG KRRAAA PPP DGSDDTND DDDH
025        --------E -------- ------- ---- P---- -G--E--- ------- ---------- ------ EDE IVFIIELN LLLL
026        --------- -------- ------- ---- G---- -K------ ------- ---------- ------ GGG IVPLLDLD SGIK
027        --------- -------- ------- ---- Q---- -I------ ------- ---------- ------ EEE KKPGASPG HKHK
028        --------- -------- ------- ---- ----- -T------ ------- ---------- ------ KNN GDNVDEE- SESR
029        --------- -------- ------- ---- ----- -K------ ------- ---------- ------ --- EEGGG-Q- QDDA
030        --------- -------- ------- ---- ----- -K------ ------- ---------- ------ --- ---SS--- ----
031        --------- -------- ------- ---- ----- -F------ ------- ---------- ------ --- -------- ----
032        --------- -------- ------- ---- ----- -A------ ------- ---------- ------ --- -------- ----
033        --------- -------- ------- ---- ----- -S------ ------- ---------- ------ --- -------- ----
034        --------- -------- ------- ---- ----- -H------ ------- ---------- ------ --- -------- ----
035        --------- -------- ------- ---- ----- -F------ ------- ---------- ------ --- -------- ----
036        --------- -------- ------- ---- ----- -W------ ------- ---------- ------ --- -------- ----
037        --------- -------- ------- ---- ----- -N------ ------- ---------- ------ --- -------- ----
038        --------- -------- ------- ---- ----- -Y------ ------- ---------- ------ --- -------- ----
039   53   TSTTNNNNE TTTTSTSS ATTTTTT TTTT GSTST TGDTKYDD ------- ---------- YYHDDD VVV APVGGEDK AAAG
040   54   GGEEGGGGS EDDDDEDE GGCCGGR GGNK QGGGG TSSNTPGN ------- TTSNSTHHSS SGDNNN KKK EEDEEPKK TVAY
041   55   NNQQRRRRK EEEEEEEE QHKQLKQ QDEE RRQHR ENKVLQTL ----VVV TTTTTTTTII LNCTTT III TTRISQMI MLMT
042   56   HHYYYYYFT LLLLLLLL EEEEKVP KLVI VIVFK KYIIKRLP DDDDPPP RKKKKKKKDD TTTLLP PPP RREKRRLH KKTT
043   57   YYYYYYYYF YYYYYYYY YYYYLWY VACC VVVVI CVVMYVYV VVVVVVV VVVVVVVVVT VVIVVV VVV VVCVVVVC VVVV
044   58   AAAAAAAAA AAAAAAAA AAAAAAA AAAV AAAAA VAAAAIAA AAAAAAA AAAAAAAAAA AAAAAA AAA AAAAAAAA AAAA
045   59   MMMMMMILM IVVVIIII AAVVAGI LIVI LMMLI VLKTVVVI VVVVIII IIVIVVVVVV VVVVVV III VIIVVIVV VVVV
```

FIG. 1. Alignment of family of homologs of protein kinase C.

```
046   60   KKKKKKKKK KKKKKKKK KKKKKKK KKKK KKKKK KKKKKKKK KKKKKKK KKKKKKKKKK KKKKKK KKK KKKKKKKK KKKK
047   61   IIIIVVVTI IIIIIIII IIIIVFM IIII KKKSE IKTEKAAH IVVVQQQ TTTTTSSTTM TTASSS EVE TTTTTSAS MMMM
048   62   LLLLLLLLL LLLLLLLL IIIIIFI IIVV IIIVI LIIVLGLV LLLLVVV LLLLLLLMLM LLLCCC LLL VVVLLLLL LLLL
049   63   DDDDKKKKK KKKKKKKK NNDDKKE NPNN RRRRK KYPRK-CE KKKKNNN KKKKKKKKRK KKKRRR RRL NNNKKRKN KKKK
050   64   KKKKKKKKK KKKKKKKK TTVIKAT KIRR LLLVT PVVLV-KK VVVVKKK PPPPPQQPEE EEEEEE EEK EEEKRKER SSPE
051   65   QQQQEEQHR DDDDDDDD KKTTQYK KRAA EEEPS VTEEK-DD VSSACCC GGGGGGGGGG DDDTTT ANS SANGGGAI TTSN
052   66   KKKKIIQTH VVVVVVVV KKGGTSY V-TS SDSNE KSQLF-GR DQQQTTT TTTTTSSSTT TTALLL TTT AAASAAST AAAA
053   67   VVVVVVVII VVVVIVII LLGGPAR L-KK EEEGF K-NDS-LI PPPPKKR MMMMMMMMMM MMMPPP SSG SSTTTSED RHHS
054   68   VVVVVVVVV IIIIVIVI SSGGKKE A-AA DSEGK K-NEG-NS TTTTNDT ---------- ---PPP PPA LMDDD-SI SALP
055   69   KKKKRRKKD QQQQQQQQ AASS--- K-FY EEEGD K-SAP-ED PAAARLL ---------- ---DDE KKE RRRQQ--G SDTS
056   70   LLLLLLMLT DDDDDDDD RRFF--- S-LL GGGGG I-T-K-KW EEEELRR ---------- ------ --- -------- ----
057   71   KKKKKKKKR DDDDDDDD --SS--- D-HH ---G- -------G ----AAA ---------- ------ --- -------- ----
058        --------- -------- --AS--- --KK ---G- -------E ------- ---------- ------ --- -------- ----
059        --------- -------- --EE--- --EQ ---G -------L ------- ---------- ------ --- -------- ----
060   72   QQQQQQQQQ DDDDDDDD DDLLDEG M-RR VVVLL KSIKEWQN QQQQSSS SMSSSSSSSS EAPLIL AAS EEEEEEAE EEEE
061   73   IIIIVVVVQ VVVVVVVV HHRRKKR QYDD PPPPD RPIFRYKG FAAARQQ PPPPVAPVTE VLLKKK NNS RRRKKFRV KKRL
062   74   EEEEEEEEE EEEEDEDE QQEEEEE GAKK SSSIM EQNRNTKT QQQQRRR EEKEQGVEAD EKHAAA KKE IITISAQS QEER
063   75   HHHHHHHHH CCCCCCCC KKAAMNV RSRR TTTSS IRQQRSLR AAAASSN AAASARPAAD EDEKKK EEE EENEEEDQ AAAD
064   76   TTTTTTTTI TTTTTTTT LLTTVIC IITT AAATA KILILTQV FFFFFFF FFFFFLFFFF FFFFFF IIF FFFFFLFF LLLL
065   77   LLLLNNNNR MMMMLMLM EELLMRE EGIV IIIVI IYVLLSAP RKKKWWW LLLLLPLLLI LLLLLL LLL LLLLLLQL MMML
066   78   NNNNDDDDS VVVVVVVI RRKKLDS RMRR RRRRR LNRMQHRM NNNNAAA QQEEEAAAQE KEAQQQ DDR NNSKKQRT SSSS
067   79   EEEEEEEEE EEEEEEEE EEEEEEE EEEE EEEEE EEEEEEEE EEEEEEE EEEEEEEEEE EEEEEE EEE EEEEEEEE EEEE
068   80   KKKKRRRRK KKKKKKKK AAVVIIL IIAA IIIVV NLLLVALV VMMMLLL AAAAAAAAAA AAAAAA AAA AAAAAAVG LLLF
069   81   RRRRLLRRQ RRRRRRRR RRDDESR SLSS SSSAK LNSESRAV AQQQNNN QQQQNNNNAK AAARKR YYY SSSHHQEI KKKN
070   82   IIIIMMMMI VVVVVVVV IIIIVIV YMLL LLLLY RLIVILLL VVVVVII VIVILLLVIV VIIIII VVI VVVLLLLI IIVV
071   83   LLLLLLLLM LLLLLLLL CCLLMML LMGG LLLLL GLVLQLHL LLLLAAA MMMMMMMMMM MMMLLL MMM MMMMMMLM MMLL
072   84   QQQQSSKSQ AAAAAAAA RRRQNNR RRQQ KKKRQ GYKHRRAK RRRRRRR KKKKKKKKKT KKKKKK AAA KKKSSSTK SSSK
073   85   AAAAIILIG LLLLLLLL LLKKQCR LLII EEERE TINKARRK KKKKLLL KKLKTQQTKK EEDQQQ SGS GEEKKNMD HHYQ
074   86   VVVVVVVVA LPPPGMGG LLVVLLV LLLL LVLLM NMVCLLVV TTTT--- LLLLLLLLFL IMLYYY VVE FFFFFFLF LLLV
075        --------- DGGGGDGE --SS--- ---- -N-E- -TK-K-SS ------- ---------- ------ --- -------- GGG-
076        --------- ---KR-R- ------- ---- -D-A- -------S ------- ---------- ------ --- -------- ----
077        --------- ---PG-G- ------- ---- -E-F- -------- ------- ---------- ------ --- -------- ----
078        --------- ---PP-P- ------- ---- -N--- -------- ------- ---------- ------ --- -------- ----
079        --------- ----G-G- ------- ---- ----- -------- ------- ---------- ------ --- -------- ----
080        --------- ----G-G- ------- ---- ----- -------- ------- ---------- ------ --- -------- ----
081   87   NNNETTESH KKK-RKRK KKGGNHR RRYY KNREQ IGPNGDSD RRRRRRH RRRKQQQQRQ KKHSSN DGE TNDNDKQS PQNN
082   88   FFFFHHHHS PPP-PPPP HHHHHHH HHHH DRHHH ISHSHHHF HHHHHHH HHHHHHHHHH HHHHHH NSH CCTHHHHH HHHH
083   89   PPPPPPPPD PPP-HPHP PSPHRPA PPPP DSPPP TSEPDPPS VVVVDDD EDDDDQPDNP PPKPPP PPV HHYPPEQP LEMP
084   90   FFFFFFFFF FFFFFFFF NNNNNKN HNHH NNNNN LRNYHAYG NNNNNNN KKKKKRRKRN NNNNNN HYN HHHNHNHN HHHH
085   91   LLLLIILII LLLLLLLL IIIILLI IIII ICIVV LVIIIIIV IIIIIII LLLLLLLLLL LLLIII VVL VVVIIIIV VIIV
086   92   VVVVIIIIV TTTTTTTV VVIIIVI ILCC VVVVI AAIVVLII LLLLVVI VVVVVVVVVV VVVVVV CSL VVVLLVVL VVVI
087   93   KKRRRRRRR QQQQQQQQ RRQQQQQ KRRR RRSRE VPTDEPTR LLPLRRR QPQQRRRKAQ QQRRRR RRK RRRKKRRS NNNN
088   94   LLLLMMMML LLLLLLLL LLLLLCL LLLL LLLLL VLFFLLLL FFFFVVV LLLLLLLLLL LLLLLL LLL LLLQLLFL LLLL
089   95   EEEEWWWWY HHHHHHHH HHKKYVV YYFF YLQMI KCYYMLHL MMMMVVV YYYYYYYHYY LILIII LLL LLLLLVFL LLLY
090   96   FFYYGGGGR SSSSSSSS DDDDADE DDEE DDDDD DDGGDDRD GGGGAAA AAAAAAAAAG GGGGGG GGA GGGGGGGG GGGG
091   97   SSASTTTTT CCCCTCTC SSTTAAV VVMM IIVVI PAAASLVW YFFFAAA VVVVVVVVVV VVVVVV IIV VVVVVIVI AAAA
092   98   FFFFFFFFF FFFFFFFF IIYYIFF IWCC VLLCF VKYFWHLF MMMMSSS VVVVVVVVCC CCCCCC CCC VVCCCCCC CCCC
093   99   KKKKQQQQK QQQQQQQQ SSEEEEE KTTT HHMAM SRYFEVEE TTTTTTT SSSSTTTTST TTTTTT LLM SSSLLFTL TTTS
094  100   DDDDDDDDD TTTTTTTT EETTTET SDLM SAQTA -VNIHVTR KRRRRRR EEEERQQKQK RRHQQQ TTS KQRLLDER KHIQ
095  101   NNNNAAASS VMMMPVPM EENNPKQ KHSS DEDSY -RQEGSEP ----TTT ----E---EH EEEKKK --- GGGNNTGS GGGD
096        --------- -------- ------- ---- A--R- --H----- ----PPP ---------- ------ --- -------E ----
097        --------- -------- ------- ---- ---T- --I----- ----AEE ---------- ------ --- -------- ----
098        --------- -------- ------- ---- ---D- --N----- ----GDG ---------- ------ --- -------- ----
099        --------- -------- ------- ---- ---R- -------- ----SSS ---------- ------ --- -------- ----
100        --------- -------- ------- ---- ---E- -------- ----NNN ---------- ------ --- -------- ----
101  102   SSSSQQRQK DDDDDDDD GGTTHAE DQNN HSSID RDNGGGDD DPPPSSS EEEEEEEEER PPAQQQ SSS QQQEEERG GGGG
102  103   NNNNQQNQY RRRRRRRR HFFFENR EHHH KKRKN TQEAFVAS NGGGLLL PPPPPPPPPP PPPPPP TTQ PPPPPPPS PPPP
103  104   LLLLIIIVL LLLLLLLL HHFFIIV IMFF LLLVL PVIVLTIF LFVFGGG IIIIIIIIII FFFIII VVM TTAQQILP IVTL
104  105   YYYYFFFFY YYYYYYYF YYFFVVY IYYY YYYTN AIIYYCYV AAAATTT YYYYYYYYYY YYYYYY QQM LLLYYSLL YLLL
105  106   MMMMMMMMM FFFEFFEF LLLLLMM MLMM LLLLL LAIMMLVL IIIIIII IIIIIIIIII IIIIII LLL VVVILLMV IVVL
106  107   VVVVIIVVL VVVVVVVV IVVVFVV VALL VVIVV IVLCQVVI VIIIIII VVVVIIIIVV IIIVVV IVI VIVIIIVV IIII
107  108   MMMMIMMMM MMMMMMMM FFFFMLM ILFF FFFFL FLMMVLLL TTTTMMM TTTTTTTTQT TTTMMM TTT MMMLLMFL TTTV
108  109   EEEE-DDDE EEEEEEEE DDDDEEE EEEE EEEEE EPEEEPQE QQQQEEE EEEEEEEEEE EEEEEE QQQ EEEEEEEP EEEE
109  110   YYYY-YYYA YYYYYYYY LLLLYML YYYY FFFHF HYYYLKYR WWWWFFF YFFYYYYFYY FFFLLL LLL LLLLLHYY YYYY
110  111   VVVVLIIIC VVVVVVVV VVMMIVA AVVV LLLVL VYSMCYCP CCCCGGG MMMMMMMMMM MMMVVV MMM MMMMMMMM CCCA
111  112   PAPPKEEEL NNNNTNTN TTKKEST GPSS DDSDP NPDDEQPE EEEEGGG SSCNAEEASK TSCQQQ PPP ATKEEERK RCCK
112  113   GGGGVGGGG GGGGGGGG GGKRGGG NDGG LMMQT NHCGNANP GGGGNNN KKHKKNNKKH YHNGGG FYL HRKGGAHH YYYY
113  114   GGGGGGGGG GGGGGGGG GGGGGGG -GGG ----- TEGGG-GV SSSSVVV GGGGGGGGGG GGGGGG GGG GGVGGGGG GGGG
114        --------- -------- ------- ---- ----- -------Q ------- ---------- ------ --- -------- ----
115  115   EEEEREEEE DDDDDDDD EEEEEEE EEQQ DDDDD DESSSDDD SSSSTTT SSSSSSSSSS NNNDDD CCC DDFDDDDD DDDS
116  116   MMMMLLLLL LLLLLLLL LLLLLLL LLLL LLLLL FFLLLLLL LLLLLLL LLLLLLLLLL LLLFFF LLL LLLLLLLL LLLL
117  117   FFFFFFFFW MMMMMMMI FFFFFFF FFLL KKKRE KRDDDYFF YYYYHHH LLLLLVVLLL LLLLLL LLL KKKLLLNR VLLR
118  118   SSSSSSSST YYYYYYYF EEDDEED DHHH RKKTV QTKKRTTD KHHHQQQ DDDDDDDDDN DDETTS DDD SSSTSSRN DNNG
119  119   HHHHLLLLI HHHHHHHQ DDYYRRR YYYY YYYYV LFIIFYYF HHHHVVV FFFFFFFFFF YFYFFF YHY YYYYYYFF YFFF
120  120   LLLLLLLLL IIIIIIII IILLIII IIII MMLLI YYLYLLII LLLLIII LLLLLLLLLL LLLLLL VVV LLLLLLLI LLLL
121  121   RRRRRRRRR QQQQQQQQ VVTTVII VRII EDDDK QRSDESTT HHHHYYY KKKKKKKKRR RRRRRR RRR RRRRRRRR HRRR
122  122   RRRRKKKKD QQQQQQQQ AAEEDDA QKQQ GRSKD TDVEEREE VVVVGDG GENDSTTSER ESRTTS EEN SSAKGASN RRRE
123  123   IIIISSSSR VVVVLVLF RRKKEEK RHHH IIIAK --YSQRKR QAAAAAA EGPGDPPDGH CATEEE HNN LLHAAAHE NKKS
124  124   GGGGKQQQG GGGGGGGG EEVVDDG DGGG PSPPS --KSGLKG EDDDATT TDEEESSEDE NGDGGG KRR RRRRRRGT KARR
125  125   RRRRDRRRS KRRRKKKK YYTTYFS KPSS KEPPI --REQNVA TTTTGRR GGGGGGGGGK RRKAAP DGD PPPMKAPH HEDK
126        --------- -------- ----H-E ---- DTGPL --FILPY- KRRRHSS KKQRGIISRT QESRRH NRK EERAQTDN TASV
127        --------- -------- ------- ---- QGQG- --VGS-Q- ----PPP YYDAKKKKYL ETL--- --- AMSTKSA- FMFM
128        --------- -------- ------- ---- PAY-- --QGR--- ----EEE ---------I ------ --- EEGFFTK- LLIP
129  126   FFFFFFFFF FFFFFFFF YYLLLLF MLIL LLMLF LLFILLGL FFFFLLL LLLLVLLQLG VLLLLL ILI PPPLLLLP LLLL
130  127   SSSSPPPPE KKKKKKKK SSSTTTT SSRK GDDPT TPNDDGNQ QDDDSSS RKRKLTNPHN NDPRRK GGS TSTTTSGT SEDT
131  128   EEEENNNND EEEEEEEE EEEEEEE EEEE APSAP DIEEERSE MMMMLLL LLLLLIVLFM AAPVMM SSS LLYLLLLV YLLM
```

FIG. 1. *Cont'd*

```
132 129  PPPPPPPPS PPPPPPPP AAKKVRR QRHH DRSEA YKLPFPHD FVVVGGG PPPPPNNPEG VVIKKK QQK QSGVTSGK TREG
133 130  HHHHVVVVT QHHHHQHV DDEEDED EEQH ILLTD EGTQRQLL QQQQKKK QQQNKKKKDL VAITTE YDA EKRDDEQD DDDD
134 131  AAAAAAAAT AAAAAAAA AATTTCA AAAA VVVII IIILVIIA LLLLCCC LLLLLLLLLL LLLLLL LLL MMILLLLL LLLL
135 132  RRRRQKKKR VVVVAVAV SSRRMIT RARR KQKKK RKSAWAKR IIIILLL VVVVILLIIL LLVLLI LLL IIYVLLLI VLLI
136 133  FFFFIFFFF FFFFFFFF HHKKVKR RHKK KKSDA YKKFKATG DDDDKKK DDDDDDDDYD YYQQQK NNN QQQDDAAG GHSS
137 134  YYYYFYYYY YYYYYYYY CCIIFYV FYFF FFYLW YYIIIVVF IVVVYYY MMMMFMMFIM MMMMMM WWW MMVLIMVF FFFF
138 135  AAAAAAAAT AAAAAAAA IIMMVML FLAA MTLMM LIAALSFF AAAASSS AAAASAASAC AAAVVM CCS AAACCCAG SSSA
139 136  AAAAAAAAA AAAAAAAA QQRRRRQ QSRR MYYRL FWYNVRLW RRRRLLL AAAAAAAATI TTSGGE VMT AGIVLISL YSYW
140 137  QQQQEEEEC EEEEEEEE QQAAQQM QQGG QQQQM EEGAEQQQ QQQQDDD QQQQQQQQQQ QQQDDN QQQ EEEDDDQQ QQQQ
141 138  IIIIVVVVV IIIIIIII IILLIIV IIII LLIFT LLVVVLLV TTTTVVI IIVVIIIIVV IIIAAA III IIIIIVVV VVVI
142 139  VVVVCCICV SAAAASAA LLLLCSL ILAA CVLLL LLLIALIL AAAAVVV AAAAAAAAAS SAAAAA AAA AAASCAAA AAAS
143 140  LLLLLLLLE IIIIIIIA EEEEDED SDSS KNQRR KRNHLSSE QQQQNNN SDEAEEEESK SSSAAA KKK DDDKKNAK NQKQ
144 141  TTTTAAAAA GGGGGGGG AAVVGGG AAAA GGGGG AAGGGAAA GGGGGGG GGGGGGGGGG AGGGGG GGG GGGGGGGA GGGG
145 142  FFFFLLLLF LLLLLLLL VVIIIVV VVLL IVILV LLLLLVVV MMMMLLL MMMMMMMMMM MMMMMM MMM MMMCCCMM MMMM
146 143  EEEEEEEEA FFFFFFFF LLCCLER EAIQ ANVDY DKDKQDER DDDDLLL AAAAAAAAET ESSEEE NSS AAAVVSVK DAAQ
147 144  YYYYYYYYY FFFFFFFF HHATFYY YHYY YFFFH YFHEFYHH YYYYFFF YYYYYFFFYY YYYYYY YYY YYYYYYYY FFFY
148 145  LLLLLLLLL LLLLLLLL CCLLMIL CCLL CCCLC CVLLIICC LLLLLLL VIMIIIIILL LLLLLL LLL LLLLLLLL LLLL
149 146  HHHHHHHHH HQQQHHHH HHHHHHH HHHH HHHHH HHYKHHHH HHHHHHH EEEEEEEEEE EEEEEE EEE NNAEEEAA AAAA
150 147  SSSSSSASS KSSSNKNT QQKKKKA RRAA SSSAR SSREHRSN AAAASSS RRRRREEQSR KSASSS DDE AAARKDGS SSSE
151 148  LLLLKKHKK RKKKQRQK MMLLMQL HFNN HRRNN MKQQKQVC KKKKQQQ MMMMKRQRKH KRRKKK RVK KNKMMMLK KKKM
152 149  DDDDDDNDG GGGGGGGG GGNNRGG KRNN RRRCF GGKNNGGG NNNNSSS NNNNNNNNQN NNHCCH RRR KKKHRHHK NNNK
153 150  LLLLIIIII IIIIIIII VVIIVII IFII IIVII IIIIYIIV IIIIIII YYYYYYYYLY FYFCCC LLL FFFFFFFF CCCL
154 151  IIIIIIITI IIIIIIIL VVVVLVT VRVV LILVL MIIIVIYL IIIIVLL VIIIIIIIII IIIIII VVV VVVIIVVV VIIV
155 152  YYYYYYYYY YYYYYYYY HHHHHHH HHHH HHHHH HHHHHHHH HHHHHHH HHHHHHHHHH HHHHHH HHH HHHHHHHH HHHH
156 153  RRRRRRRRR RRRRRRRR RRRRLLR RRRR RRRRR RRRRLRRR RRRRLLL RRRRRRRRRR RRRRRR RRR RRRRRRRR RRRR
157 154  DDDDDDDDD DDDDDDDD DDDDDDD DDDD DDDDD DDDDDDDD DDDDDDD DDDDDDDDDD DDDDDD DDD DDDDDDDD DDDD
158 155  LLLLLLLLL LLLLLLLL LLLLLLL LLLL LLLLL VIIVLILI MLLLLLL LLLLLLLLLL LLLLLL LLL LLLLLLLL LLLL
159 156  KKKKKKKKK KKKKKKKK KKKKKKK KKKK KKKKK KKKKKKKK KKKKKKK RRRRRRRRAA AAAAAA AAA AAAAAAAA AAAA
160 157  PPPPPPPPP LLLLLLLL PPPPPPP PLII PPPPP PPPPPTPD SSSSPPP AAASAAAAAA AAAAAA AAA AAAAACTA AAAA
161 158  EEEEEEEEE DDDDDDDD EEEEEEE EEEE QQQEN HTSTAEEE NNNNAAA AAAAAAAARR RRRRRR RRR RRRRRRRR RRRR
162 159  NNNNNNNNN NNNNNNNN NNNNNNN NNNN NNNNN NNNNNNNN NNNNNNN NNNNNNNNNN NNNNNN NNN NNNNNNNN NNNN
163 160  LLLLIIIIL VVVVVVVV LLIIIIL LIII LLLIL VFVIVIII IIIIIII IIIIVIIIVC CCCCCC VVV CCCCCCCC VVII
164 161  LLLLLLLLI MMMMMMML LLLLLML LLMM LLLLL MLLLMFML FFFFLLL LLLLLLLLLL LLLLLL LLL MMMLLLLM LLLL
165 162  IIIILLLLL LLLLLLLL LLLLCCY LIII IIIVF IFICIIVI LLLLIII VVVVVVVVIV VVVVVV VVV VVVVVVVL ILLV
166 163  DDDDDDDDD DDDDDDDD AADDVVY DKSS NDDTS DNNSTNGD HHHHSSS GGGGSSSSGG GGSTTT KKR AAASSTGD CTTA
167 164  QQHHKKRKH SSSSASAA SSDDNNH EVDS KKDSP HLSAFTNL EEEEEEE EEENEDDAES EDEEEE TSL HEDVEEQE ENHE
168 165  QQQQNNNNR EEEEEEED KKDNTKP HNSS DEKGD EEKNEPDS GGGGQQK NNRGSTTSNE NNHKKK PPL DDDKKSGK GGGG
169 166  GGGGGGGGG GGGGGGGG LLMMTTG LESG GGGGG NLGQGEGR LLLLDDD LLLLLLLLNN HKNNNN QN- FFLDQTLF KHRR
170      --------- -------- KK--GGT -Q-- ----- RG-G--NG ------- ---------- ------ --- ---YYG-- ----
171      --------- -------- GG--HTD ---- ----- -------- ------- ---------- ------ --- ---TGS-- ----
172      --------- -------- AA----S ---- ----- -------- ------- ---------- ------ --- ---SST-- ----
173      --------- -------- ------- ---- ----- -------- ------- ---------- ------ --- ---PCD-- ----
174      --------- -------- ------- ---- ----- -------- ------- ---------- ------ --- ---RSR-- ----
175      --------- -------- ------- ---- ----- -------- ------- ---------- ------ --- ----RR-- ----
176      --------- -------- ------- ---- ----- -------- ------- ---------- ------ --- -----R-- ----
177 167  YYYYHHHHY HHHHHHHH AANNLSK NQEE NNTTQ KRQTTDTE TTTTVVV VVAIMSSVVV LLIVVT HHA TTTIVTVT LVIK
178 168  IIIIIIIIA IIIIIIIV VVIIVII VIII LLIVI LGIVLIVI VVVVCCC CCCCCCCCAV VVVLLL VVG VVVVVVVV VATM
179 169  QQQQKKKKK KKKKKKKK KKKKKKI KKKK KKKKK RVKKKCYK KKKKKKK KKKKKKKKKK KKKKKK KKE KKKKKKKK KKKK
180 170  VVVVIIIIL IIIIIIII LLLLILI IIII LLLLV LLLLILLL IIIIIII VIIIIIIIIV VVIIII IID IIIIIIIV IIII
181 171  TTTTTTTTV AAAATATA AATTIIT AAII GAAAA IVCCGGAI GGGGSSS AAAAAAAACA AAASSS TTH GGGGGGGA CGCS
182 172  DDDDDDDDD DDDDDDDD DDDDDDD DDDD DDDDD DDDDDDDD DDDDDDD DDDDDDDDDD DDDDDD DDD DDDDDDDD DDDD
183 173  FFFFFFFFF FFFFFFFF FFFFFFF FFFF FFFFF WFFFFFFF FFFFFFF FFFFFFFFFF FFFFFF FFF FFFFFFFF FFFF
184 174  GGGGGGGGG GGGGGGGG GGGGGGG GGGG GGGGG GGGGGGGG GGGGGGG GGGGGGGGGG GGGGGG GGG GGGGGGGG GGGG
185 175  FFFFFFFFF MMMMMMMM LLFFLLL LMLL LLLLL LLVVMALS LLLLCCC LLLLLLLLLL LLLMMM LLL MMMLLLML LLLL
186 176  AAAAAAAAA CCCCCCCC AASSAAA SASS AAAAA AASSAAAG AAAASSS AAAAAAAAAA SAASSS AAA TTTAAASA AAAS
187 177  KKKKKKKKK KKKKKKKK IICCRRS NTNN RRRRR EEKGSCTA TTTTEQQ RRRRRRRRRR RRRRRR KRK RRRRRRRR RRRR
188 178  RRRRYYEYK EEEEEEEE EEQQRRA IVII ASAIA FAKNVFTL VVVVKKK LLLLVLLVVY LLFEEQ LLL DDDDDDDD DDDD
189 179  VVVVVVVVI HNNNNHNN VVLLYLR MEYF FFFYI YQLLWVEL KKKKLLL IIIIIIIIIV MMMEEE LLL IIIIIIIM IIIV
190 180  KKKKPPQPG MIIIVMVI EQDQNEK TPDD GGGSP HMIVPQPK STTTEQQ EEKEEEEEAL TRKAAE GDS YYYYYYYY MMKY
191 181  GGGGDDTDF MWWWFMFV GGPPPSK DNSY VVIYA PDNAVGYD RRRRDDD DDDDDDDDDD GDEDDD AIS EEEKKKSD RNNE
192 182  RRRRVVVVG DDDDPDPG EDGGNAG GDRR PPPQP GYSSPSST WWWWLLL NNDNNNNNDD DDDGGG EDD TTTNNSTK DDDE
193 183  TTTTTTTTK GGGGGGGD QQEEEGD NSKK LLIMH QKILRRSV SSSSLRR EEEEEEEEEQ TTTVIV EES DDDDDDDE SSSD
194      --------L VVVVSVSK QQKKKSD FCQQ RRRAE ESAAGS-Y GGGGCGC YYYYYYYYYY YYYYYY KTN YYYYYYYY NNNS
195      --------T TTTTTTTT AALLLLC LLLL ANVLI YM---S-- SAAAFRR TTNTTTTTCT TTTAAA EEE YYYYYYYY YYYY
196      --------- TTTTTTTT WW----L ---- YYY-L -I---P-- QQQQQQP AAPAAAAAPS AAAAAS YYY RRRRRRRS IIVV
197      --------- -------- ------M ---- ----- -S---F-- QPPPTA- RRCRRRRRKS HHHSST HHK KKKKKKVV SVVK
198 184  WWWWYYWYW RKKKRRRK FFRRKKK KEHH TTTTT NADKMYDT EEEEYPH QQQQEEEEQG AAAGGG EDA GGTGGGGT GGGS
199 185  TTTTTTTTT TTTTTTTT GGEEVVT TNTT HHHPS VNTTEGFD QQQQPHH GGGGGGGGGG GGGLLM GGG KKKEEERG SNNQ
200 186  LLLLLLLLF FTFFFFFF FFVVNLT SYFF EEEVN RRFNRIGF PPPPLII AASAAAAAST AAARRK GGG GGGGGGTA TAAG
201 187  CCCCCCCCC CCCCCCCC AACCFFC CCCC IIVVV VAVIEACD TSSSGGG KKKKKKKKRK KKKQQQ KKK LLLLLLMK YRRR
202 188  GGGGGGGGG GGGGGGGG GGGGGGG GGGG VVVVV AGGGGGGG GGGGGGG FFFFFFFFFF FFFVVI VVM LLLLLLLL LLLI
203 189  TTTTTTTTT TTTTTTTT TTTTTTT SSSS TTTTT STTCDTST SSPSTTT PPPPPPPPPP PPPPPP PPP PPPPPPPP PPPP
204 190  PPPPPPPPP PPPPPPPP PPPPPPP PLLL LLLLR RRSQCILR VVVVYYY IIIIIIIIVI IIIVVV III VVVVVVIV LVVV
205 191  EEEEDDDDE DDDDDDDD GGSSEEE NHYY WWWWW YGTSEDFV LLLLTTT KKKKKKKKKK KKKKKK KKK RRRRRRRK KKKK
206 192  YYYYYYYYY YYYYYYYY YYYYFFY YYFF YYYYY FFYYYTYY WWWWHHH WWWWWWWWWW WWWWWW WWW WWWWWWWW WWWW
207 193  LLLLIIIIV IIIIIIII LLLLLVI ALAA RRRRR KRMMINMS MMMMRQQ TTTTTTTTTA TTTTTT MML MMMMMMMM MMMM
208 194  AAAAAAAAA AAAAAAAA SSAASAA AAAA AASAA GASAAASP AAAAAAA AAAAAAAAAP AAAAAA AAA ASPAASPA AAAA
209 195  PPPPPPPPP PPPPPPPP PPPPPPP PPPP PPPPP PPPPPPPP PAAAPPP PPPPPPPPPP PPPPPP LLL PPPPPPPL PPPI
210 196  EEEEEEEEE EEEEEEEE EEEEEEE EEEE EEEEE EEEEEEEE EEEEEEE EEEEEEEEEE EEEEEE EEE EEEEEEEE EEEE
211 197  IIIIVVVVI IIIIIIII VVIIVVV VILL VVVVL LVRRVVCW VVVVLIL AAAAAAAAAV SGGAAA SSC SSSSSSSS SSSS
212 198  IIIIVVIVI IIIIIIII LLIIVIL IVLL LLLLL LLIILLQI IITILLL AAAAIIIIIL LLLLLL III LLLLLLIL IIIL
213      --------- -------- --EQ--- ---- LLLLF VM-K--RR RRRR--- ---------- ------ --- -------- ----
214      --------- -------- --CC--- ---- ----- ---S--E- MMMM--- ---------- ------ --- -------- ----
215      --------- -------- --SS--- ---- ----- ---L--V- QQQQ--- ---------- ------ --- -------- ----
216      --------- -------- --MM--- ---- ----- ---N--K- ------- ---------- ------ --- -------- ----
217 199  LLLLSSTSL AAAAAAAL RRNGNNV SSKK GGGQG DKQDAASY DDDDKKK LLLLNNNNIN AAANNN LLR KKRMIVLQ FFFF
```

FIG. 1. *Cont'd*

```
218 200  SSSSTTTTN YYYYYYYY KKHHYYR GHAA GSSSA YCGRNGSH NPPPGGG YYFYFYYFYY YYFYYY HRN DDDDDDYT NDND
219 201  KKKKKKKKK QQQQQQQQ DEPPDEK KKNQ KRATK QGNAHDSR NNNNEEE GGGGGGGGGT NNNGGG RRR GGGGGGRQ SCCH
220 202  GGGGPPPPG PPPPPPPP PAGGQPP LPPP QHR-H MAVTLPFY PPPPGII RRRRCTTSKR KKTRRR IRV VVVIVLKK LVVI
221 203  YYYYYYYYH YYYYYYYY YYYYIIY YYYY YYYYY YQYYYYAH FYYYVAA FFFFFFFFFF FFFYYY YFF FFYFFFFF YYYY
222 204  NNNNNNNND GGGDGGDG GGGGSGT ARTT SSSAT DSSTDTTG SSSSTTT TTTTTTTTSS SSSSSS TTT TTSTTTTT TTTT
223 205  KKKKKKKKI KKKKKKKK KKKKDYN GGGG TTTTS YTIVKTAR FFFFPPP IIIIIIIIIS ITSSSS HHS TTSTNTTT TVFT
224      --------- -------- ------- PAPP ----- -------- ------- ---------- ------ --- -------- ----
225 206  AAAASSSSS SSSSSSSS PPEEKES EPEE GGPPA SKKQPTPS QQQQKKK KKKKKKKKKK KKKEEE QQK SYAQHQEK LQEQ
226 207  VVVVIIVVA VVVVVVVV VVVVTTV VAVV VVVVI LIGSAVNA SSSSAAA SSSSSSSSSS SSSSSS SSS SSSSSSSS SSSS
227 208  DDDDDDDDD DDDDDDDD DDDDDDD DDDD DDDDD DDDDDDDA DDDDDDD DDDDDDDDDD DDDDDD DDD DDDDDDDD DDDD
228 209  WWWWWWWWY WWWWWWWW LIMMMMM VVVI IIIMI MIVIIIVV VVVVIII VVVVVVVVVV VVVVVV VVV MVVVVVVV VVVV
229 210  WWWWWWWWW WWWWWWWW WWWWWWW WWWW WWWWW WWWWWWWW YYYYYYY WWWWWWWWWW WWWWWW WWW WWFWMWWW WWWW
230 211  AAAASSSSS AAASSAAA AASSSSA SSSS SSSSS SSSSSSAS SAAASSS SSSSSSSSSA AAASSS SSA SSSSAASS SSSS
231 212  LLLLFFLFL YFFFFYFY CCTTLIL CCFF IVIVV LVLLLALL YYYYFFF FFFFFFFFFF FFFFFF YYF FFFFFFFF FYYF
232 213  GGGGGGGGG GGGGGGGG GGGGGGG GGGG GGGGG GGGGGGGG GGGGAGG GGGGGGGGGG GGGGGG GGG GGGGGGGG GGGG
233 214  VVVVIIVVI VVVVVVVV VVVVVVV VVVI CCTCV CVLLILII IVVVIII IIIIIIIIIV VVVIII VVV VVVIVVVV IIIV
234 215  LLLLLLLLL LLLLLLLL IIIIIII IIVV IIIII MIMSTVIL VVVVTTT LLLLLLLLLL LLLLLL TTT VVVLLLVV LLFL
235 216  IIIIIIIIM LLLLLLLL LLMMTCA LLLL FFFFF LLIIVILL LLLLLLL LQLLLLLLLM LLLLLL VVI LLLIVCLL LLLL
236 217  YYYYYYYYY YYYYYYYY YYYYYYY YYFY AAAAA ALILFFIY YYYYWWW TTTTYTTMMW WWWWWW WWW WWWWWWWW WWWW
237 218  EEEEEEEEE EEEEEEEE IITTMII VSVV EEEEE SSEEEEND EEEEQQQ EEEEEEEEEE EEEEEE EEE EEEEEEEE EEEE
238 219  MMMMMMMML MMMMMMMM LLLLLLL MLLL MMLML MLLMATLM LLLLMMM LLLLIIIILI IIITTA LLL IIMITIIL IILI
239 220  AAAALLLLL LLLLLLLL LLLLLVL LLVV CIAFM ILVAAACV MMMMTTT TVIVVVVVFF AAAFFF MML TAALLLFM FFFV
240 221  AAAAAAAAT AAAAAAAV VVAASSS CSCC NRTRL FGTLAVCC TTTTTTT TTTTTTTTTT TTTSSS TTT STTTTTTT TSST
241      --------- -------- ------- ---- ----- -------- ------- KKKKYHHYYC YYYLLL FFF LLLLLLYR LLLL
242 222  GGGGGGGGG GGGGGGGG GGGGGGG RNGG RRKRR RRGGNHKG GGGGKRR GGGGGGGGGG GGGGGG GGG AAAGGGGG GGGG
243 223  YYYYYYYYS QQQQQQQQ YYSSLLT RKKK KSKKI KRERINRD ESSSQEE RRRRKRRRQK MMMAAA SAQ EEAHQQKA GLSG
244 224  PPPPTTTTP PAAAPPPP PPPPSSM LLVV PPPPP FFFYVANI LLLLAVV VVIVIIIIVM SSASSV KKR QQQQQQQP TNSN
245 225  PPPPPPPPP PPPPPPPP PPPPPPP PPPP ILLLY PPPPLSPP PPPPPPP PPPPPPPPPP PPPPPP PPP PPPPPPPP PPPP
246 226  FFFFFFFFF FFFFFFFF FFFFFFF FFFF FFFFL FMLYPLWF YYYYYYY YYYYYYYYYY YYYYYY YYH YYYYYYWY YYYY
247 227  FFFFYYYYS DEEEDDDD WWWWLME DGDD SPHCP FFGPDFK- SSSSSSS PPPPPPPPPG PPPPPA DDE QQQPPAYP PPPP
248 228  AAAADDDNG GGGGGGGG DDHHGGD DGDD GGGGG HQGPNSR- HHHH--- GGGGGGGGGR GAGNNN GGN GGGAGAQD EGGG
249      --------- -------- ------- ---- DDDNQ GSH-GAA- IIII--- ---------- ------ --- -------- ----
250      --------- -------- ------- ---- SSSSN HLN--PC- ------- ---------- ------ --- -------- ----
251      --------- -------- ------- ---- EEEED D-----R-- ------- ---------- ------ --- -------- ----
252      --------- -------- ------- ---- IIIAV N-----G-- ------- ---------- ------ --- -------- ----
253      --------- -------- ------- ---- ----- Y-----P-- ------- ---------- ------ --- -------- ----
254      --------- -------- ------- ---- ----- ------K-- ------- ---------- ------ --- -------- ----
255      --------- -------- ------- ---- ----- ------R-- ------- ---------- ------ --- -------- ----
256      --------- -------- ------- ---- ----- ------G-- ------- ---------- ------ --- -------- ----
257      --------- -------- ------- ---- ----- ------P-- ------- ---------- ------ --- -------- ----
258      --------- -------- ------- ---- ----- ------C-- ------- ---------- ------ --- -------- ----
259      --------- -------- ------- ---- ----- ------D-- ------- ---------- ------ --- -------- ----
260 229  DDDDSSTSP EEEEEEEE EERRDDD EQEE DDDDD DDDEQSSE NGGGGGG MMMMRMMMML IIVLLL III LLLHLRLV PLPI
261 230  QQQQNNTND DDDDDDDD DDKKDNN SNNN QEQQQ QDTTSQQH NCCSEEE VVNNTTTSHK DDESSS PPP SSSSSNSN MVVP
262 231  PPPPTTPTP EEEEEEEE QQQQDDR ITSS IILLM LAPYWITD RRRRRPP NNKNNNNNSN RLLNNN AAA NNNNNNNT NNDP
263 232  IIIIMMMMM DDDDEDEE HHMMTNT PDSS FFFGE VDDDQTDE DDDDQQQ RRRRAPPPRT STSQQQ SRK EEELIFTF DSSE
264 233  QQQQKKKKK EEEEEEEE RKLLEER VVVI KKRKV RSGNKRGE QQQQHYH EEEEDEEEEE QDNQQQ EED QQQDEEED QKKR
265 234  IIIITTTTT LLLLLLLL LLMMTTL LILL IIIIT ILIILITI IIIIIVV VVVVVVVVVV VVVTTT III VVVVVVAI FFFL
266 235  YYYYYYYYY FFFFFFFF YYLLLLY FYHH FFFFF ALLFRIYI IIIILQQ LLLLMIIIIV YYYRRR SPP LLLLLLIT YYYF
267 236  EEEEEEEEN QQQQQQQA QQRRNAR KNEE RQRDR KEDSSRRK FFFFYYY DEEETQQREE EHGEEE SDD KRRNHADV NKKN
268 237  KKKKKKKNI SSSSASAA QQMMNNQ NKKK VVALA VLLQGQSG MMMMAAA QQQQANNANR LKLFFA ILL FFYYHHCY ALML
269 238  IIIIIIIII IIIIIIII IIIIVVI IIII LLLIL LCLLDAYQ VVVVVVV VVVVLLLLIV LLLVVI LLI VVVVVVIL IVIL
270 239  VVVVLLLLL MMMMMMMT KKMMLTL SRKK GGGGG GTQSLQVV GGGGVVV EEEESEEEEQ EDEEEE EEE MMIQRKTL KKKK
271 240  SSSSNNQNR EEEEEEED AASDSSR NHQK TTTLT TIRASVHF RRRRAAA RRQRQRRRRR KKNKKQ KKV DEDTSEQQ RDET
272 241  GGGGAAGAG HHHHQHQH GGGGGAG GGGG PPPPP EFIIDHNF GGGGYYY GGGGGGGGGG DGGGGG GGG GGGGGGGG GGGG
273      --------I -------- ------- ---- NNNPT EGVVAVPR YYYYDNN YYYYYYYYFI YYFGGV EEL GGGGGGRR YYFH
274 242  KKKKEEKED NNNNTNTN AANKNTK VAKK EENED LWNDPDSQ ALLLLLL RRHRRRRRRI RRRRRR RRK YLVRRRER RQRR
275 243  VVVVLLVLI VVVVVVVV YYYYWWY YYVV AEEDR YKEGRETT SSSSRRR MMMMMMMMML MMMLLL LLL LIMLLLLL MMMM
276 244  RRRRRRVKE SAAATSTS DDQQYTS TDED IVVDD AEPPLFL- PPPPPPP PPPPPVVPPE KEDPPE PPE DDEEEQEL AALE
277 245  FFFFFFYFF YYYYYYYY FFFFFFY LLYY WWWWW YLSPSSL- DDDDSSH CCCCRRRRKK RRGCCP QQQ QKRPSQRQ QQSR
278 246  PPPPPPPPP PPPPPPPP PPGGDDS PPPP PPPPP LRPPSPS- LLLLLLW PPPPVPPPPP PPPPPP PPP PPPPPPPP PPPP
279 247  SSSSPPPPK KKKKKKKK SSSSEDG KSQS DGERE DKRRTHI- SSSSSAQ PQPQEDDETK EPQEEE PPE DDERNPRE AAED
280 248  HHHHFFYFK SSSSSSSS PPPPEEE FSHH IVVDV KCLLDPL- KKKKAG- EGGDNNNNNS GGGLLQ III NNNNNMAY HFHN
281      --------- -------- EEEETAP ---- VTEVS YAPPNEP- LIIIAAA --------H- ------ --- -------- ----
282      --------- -------- WWWWFFW ---- YLSSS NAKSGS-- YSSFVVV --------Y- ------ --- -------- ----
283      --------- -------- DDDDEDP ---- LLLLF ILDDSR-- KSSSFFF ---------- ------ --- -------- ----
284      --------- -------- TTDDAES ---- PQQPM DHRKSL-- NNNNETT ---------- ------ --- -------- ----
285      --------- -------- ------- ---- DDDRT LGI-LT-- ----DAA ---------- ------ --- -------- ----
286 249  FFFFFFFFI LMMMLLLL VVYYVIV LILL LGLMA VDYFGMIV CCCCPTT CCCCCCCCFC CCCCCC CCC CCCCCCCC AAAC
287 250  SSSSNNQHA SSSSSSSS TTSSSSS SSSS DEDES SHSSQDSS PPPPGGG PPPPPPPPPA PPPPPP TTS PPPPPTPP SPPS
288 251  SSSSEEPPK KKKKRKRK PPDDDDN PSII PEEEE PYKSGIRS KKKKQKK EEAIDEEEDK EPPDDE IIL EDDDDEPD DKAE
289 252  DDDDDDDDN EEEEEEEE EETTEDL GAEE RDNSY EWEDGDEE AAAARTT SSSSEEEENE KESAAD DDD RMFDDKEP ENEE
290 253  LLLLVVVAA AAAAAAAA AAVVAAA AAVV GAGGA ACMALVLC MMMMLLL LLLLLLLLII VVVVVV VVI VLLLILVL IIMM
291 254  KKKKKKVQA VVVVVVVK KKKKKKK AQII IILAL LFTQDENQ KRRRGQQ HHYHYYYYYY YYYFFY YYY TFHWRYYY YYYY
292 255  DDDDDDDDN SAAAASAE DNDDDDD GDSS DEDQD DQDDRYSH RRRRDNN DEEEDQHNQD EDRRRR MMC DEKNDSAE ESDR
293 256  LLLLLLLLL IIIIIIIA LLLLFFF LLLL LLLLF FVFFVLLL LLLLVIN LLALILLILV LLLLLL IIT LLLLLLIV IIIL
294 257  LLLLLLLLI CCCCCCCC IIVVVII ILLL LLLLM LLVVVVLI VLLLIIV MMMMMMMMLM MMMMMM MML MMMMMLMM MMMM
295 258  RRRRSSSKK KKKKKKKK NNSSSSD KHST DSSLC DENSECNK ASSTQQQ CNEIKRMMLK RRLEEQ VVL RRQTTLRL QQKL
296 259  NNNNRRKKK GGGGGGGG KQRRNNR RRKR KAKEG KQRLWKRW DDDDRSS QLQHMLLRQL AQQQQR KKS MMRQRLGK KATQ
297 260  LLLLLLLLL LLLLFLFF MMFFLLL MMMM LMMMM LCCCMAIC CCCCCCC CCTCCCCCCC CCCCCC CCC CCCCCCCC CCCC
298 261  LLLLIIIIC MIMMLMLL LLLLILL LLLI LLLLL LFCLLLFL VLLLWWW WWWWWWWWWW WWWWWW WWW WWWWWWWW WWWW
299 262  QQQQTTTTR TTTTTTTT TTVVVKT IDVV AVITT REIQSTDS KKKKREE RKRKKKKKDS QQNAAE MMH QQHAARQH EADK
300 263  VVVVRRARD KKKKKKKK IIVVKKV VVVV YYYFM YMKKPFRL KFFFPAA KKLKEEENAH WWWYYY IIL FYHQQTRP ELAQ
301 264  DDDDDDDDN HHHHHHHQ NNQQEDD NNDD DDDNN DDNIEDNR VQQQSRR EDDDKRRRVG NDSEED DDD NNREDDEK KEDE
302 265  LLLLLLLLP PPPPPPPP PPPPQMP PPPP PPPPP HPEPPGPP KRRRAAA PPPPPPPPPP PAPPPP ASA PPSPPPPA FPPP
303 266  TTTTSSTSS GGGGAAGN SAQQGKG LSKL IAAHQ VQRERAKS EEEEALL EDEEEEEEEE STSGGR DEA NKSDHWQE ETLD
```

FIG. 1. *Cont'd*

304	267	KKKKQQREE	KKKKKKKK	KKKDASA	NTRR	NHKKK	DKERNLTD	EEEEQQQ	EEEEEDDEKE	DDDQQR	SCM	MMAQNEQM	THKK
305	268	RRRRRRRRR	RRRRRRRR	RRRRRRR	RRRR	RRRRR	RRRRRRRR	RRRRRRR	RRRRRRRRRR	RRRRRR	RRR	RRRRRRRR	RRRR
306	269	FFFFLLILL	LLLLLLLL	IIYCMLM	IIAA	IIIIW	LSSPPPIP	PPPPPPP	PPPPPPPPPP	PPPPPP	PPP	PPPPPPPP	PPPP
307	270	GGGGGGGGG	GGGGGGGG	TTTSSNT	STTT	SSSST	TSSTTSTS	LLLLSGG	TTTTTTTTTA	STRSSS	KRT	TSSTTSSS	PTTV
308	271	NNNNNNNNN	CCCCSCSC	AAAAACA	IILL	AAGAA	AAIYIALF	FFFFAAA	FFFFFFFFFF	FFFFFF	FFF	FFFFFFIF	FFFF
309	272	LLLLLLLLL	GGGGGGGG	AHEEATL	HPKK	RKKFV	REHADAPE	PPPPREE	EEEEDDDEER	AKRSSG	RRK	LLLHFRKS	SQKA
310	273	KKKKQQQQK	PPPPPPPS	EEEEQQQ	EEQN	RRMRQ	EDEAQEEE	QQQQLLL	YYYYYYYYFV	ESDTAA	EEQ	EEDRYRDE	QQQD
311	274	DDNNNNDNN	EEEEDEDS	AAAACCA	IFVV	AAAAC	ALLLILLI	IIIILLL	LILLLLLILL	IIIIIV	LLL	IIIIICVL	LIII
312	275	GGGGGGGGG	GGGGGGGG	LLLLLLL	MFVV	ALLLL	MLLTLLSR	LLLLLQQ	QQQQQRRQNM	HHHYYH	IVT	VIIQQYHV	VCVS
313	276	VVVVTTSSV	EEEEEEEE	KKAAAQR	QSEE	IQNQE	AKHEACTN	SAAAVRK	ASSSSSSSHD	QHFQQQ	IST	NSADHNAS	LSQK
314	277	NNSSEEREK	RRRRPRPE	-------	----	-----	----T---	STTTDDD	FFFFVVVVYQ	AANEED	EEV	LSYQKTRR	LFLD
315	278	DDDDDDDDD	DDDDTDTD	-------	----	-----	----D---	IIIILLL	LLLLLLLLFL	FLLLLL	FFF	LILLLLLI	LLIL
316	279	IIIIVVIVI	VIIIIVIV	-------	----	-----	--------	-------	----------	------	---	--------	----
317	280	KKKKKKKKQ	RKKKRRRR	-------	----	-----	--------	-------	----------	------	---	--------	----
318	281	NNTTNNANK	EEDEAEAL	-------	----	-----	--------	---E---	----------	------	---	--------	----
319	282	HHHHHHHHH	HHHHHHHH	HHHHHHH	DHHH	HQHHS	HTHHELLH	EEEETKK	EEEEDEDDEA	EEEQQI	SSA	KKEQQHQS	EQEE
320	283	KKKKPPPPK	AAAAGAGP	PPPPPPP	DPHP	PNPSD	PPDPVPVP	LLLLSAA	DDDDDDDDSL	THNSSA	KRE	DEPLEAAA	REKK
321	284	WWWWWWWWW	FFFFFFFF	WWFFWWW	WFWW	YYYYY	YFLWCLSW	LLLLLFF	YYYYFFFFFV	MMLIII	MMF	DEQFIILI	LQQM
322	285	FFFFFFFFF	FFFFFFFS	IVFFLLV	FLMM	FLFLF	FFILWFNM	QQQQKRR	FFFFYFFYSA	FFIRRR	AAA	LMCRRSAF	LAIM
323	286	AAAAKKSNE	RRRRRRRR	SCQQNQV	KMVN	QRNHK	LNMVVQCQ	HRRRAGG	TTTTTTTTTQ	QQSKKK	RRR	HEPNHTQS	GQSV
324	287	TTTTEEEEG	RYYYWRNR	HQQENKS	VGRR	EDDKE	PEKKEQKG	SSSSEAA	SASAAAAATT	EVSRRR	DDD	PPNFSDAT	EEEK
325	288	TTTTVVVVF	IIIIIIII	RRYYLDM	DCGG	SFLDL	ILYYMKND	LLLLLLL	TTATTTTTSL	SGNHHH	PPP	SGSFPLPF	GDSR

FIG. 1. *Cont'd*

FIG. 1. The alignment of the sequences of the catalytic subunits of 79 protein kinase homologs, derived from the alignment published by Hanks *et al.* (1). The first column contains the position numbers in the alignments, and all references in the text refer to these numbers. The second column contains the sequence number of the β-catalytic chain from porcine kidney EC 2.7.1.37, presumed to be the protein for which a crystal structure will shortly become available. The remaining 12 columns arrange sequences vertically according to functional subfamilies designated below.

Functional subfamilies 1–7: Protein serine/threonine kinase homologues.

Functional subfamily 1: Cyclic nucleotide dependent subfamily (MPI = 50%). cAMP dependent protein kinases: bovine (α form), mouse (α form), bovine (β form), mouse (β form), *Saccharomyces cerevisiae* (RAS suppressor), *S. cerevisiae* (type 1), *S. cerevisiae* (type 2), *S. cerevisiae* (type 3); cGMP dependent protein kinase: bovine. *Functional subfamily 2*: Calcium-phospholipid dependent subfamily (MPI = 80%). Protein kinase C: bovine (α form), bovine (β form), rabbit (αβ forms), rat, bovine (γ form), rabbit (γ form), rat, *Drosophila melanogaster* (related gene product). *Functional subfamily 3*: Calcium–calmodulin dependent subfamily (MPI = 45%). Protein kinases type II: rat (α subunit), rat (β subunit); phosphorylase kinases: rabbit (γ subunit), mouse; myosin light chain kinases: rabbit (skeletal muscle), chicken (smooth muscle); Putative protein serine kinase: human. *Functional subfamily 4*: SNF1 subfamily (as in alignment by Hanks *et al.* (1988)) (MPI = 50%). *Functional subfamily 5*: CDC28-cdc2+ subfamily (as in alignment by Hanks *et al.* (1988)) (MPI = 45%). *Functional subfamily 6*: Casein kinase subfamily, STE7 subfamily, and family members with no close relatives (as in alignment by Hanks *et al.* (1988)) (MPI = 30%). *Functional subfamily 7*: Raf-Mos proto-oncogene subfamily (MPI = 35%) cellular homologs of oncogene products: human (Raf), human (A-Raf), human (PKS), mouse (A-Raf), human (Mos), mouse (Mos), rat (Mos).

Functional subfamilies 8–12: Protein tyrosine kinase homologues.

Functional subfamily 8: Src subfamily (MPI = 60%). Cellular homologues of viral oncogene products: human (Src), human (Yes), human (Fgr); putative protein tyrosine kinases: human (FYN), human (LYN); lymphoid cell protein tyrosine kinases: human, mouse; hematopoietic cell putative protein tyrosine kinase: human; gene products related to Src: *D. melanogaster* (Dsrc64), *D. melanogaster* (Dsrc28). *Functional subfamily 9*: Abl subfamily (MPI = 50%). Cellular homologs of viral oncogene product: human (Abl); gene products related to Abl: *D. melanogaster, C. elegans*; cellular homologs of viral oncogene products: human (Fes/Fps), feline (Fes/Fps), chicken (Fes/Fps). *Functional subfamily 10*: Epidermal growth factor receptor subfamily (as in alignment by Hanks *et al.* (1988) (MPI = 65%). *Functional subfamily 11*: Insulin receptor subfamily (MPI = 50%). Insulin receptor: human; insulin-like growth factor 1 receptor: human; gene product related to INS.R: *D. melanogaster*; cellular homolog of viral oncogene products: human (Ros), chicken (Ros); gene product of the 'sevenless gene': *D. melanogaster*; oncogene products: human (TRK), human (MET). *Functional subfamily 12*: Platelet derived growth factor receptor subfamily (as in alignment by Hanks *et al.* (1988)) (MPI = 55%).

number of sequences that are needed to model known structures, implying that a satisfactory number of structural assignments can be made with the most reliable versions of the algorithms that we have developed. Indeed, the large number of alignable sequences has proven to be a disadvantage in our hands. Most of our computer programs were originally written to handle alignments of fewer than 25 proteins; these have had to be rewritten to handle the alignment of protein kinase catalytic domains.

Further, no crystal structure is as yet available for any member of the family. However, crystals of the catalytic domain of a cAMP-dependent protein kinase were reported in 1985 (12).

Attempting to predict an unknown structure places our approach at maximum possible risk, and we appreciate the extent to which this manuscript will (and should) be viewed as a test of the method and our ability to apply it. One obvious disadvantage with a true prediction is, of course, that the timing must be nearly perfect. A structure prediction made far in advance of a crystal structure is uninteresting because it cannot be critically tested. A structural prediction that follows, even shortly, a crystallographic model of a protein, is also uninteresting, as it might have been influenced by the experimentally determined structure.

We have recently learned that crystallographic work on the catalytic domain of the cAMP-dependent protein kinase has progressed rapidly, so rapidly in fact that a chain tracing for the protein may be available before the end of the year (Susan Taylor, personal communication, July 20, 1990). Therefore, we have hastened to document here the predictions that the method makes with respect to the folded structure of the protein kinase family. In some sense, this documentation is premature; we have not yet been able to complete an analysis that might provide a convincing argument for a single assemblage of secondary structural elements into a unique tertiary structure.

Nevertheless, we are able at this time to suggest a small set of supersecondary structures and arrangements of these into a globular form using our method, and these can be compared with the experimentally determined structure when it becomes available.

Further, the timing has forced us to make several compromises. In particular, the absence of a full set of computer programs for handling a large alignment has made it difficult to explore all aspects of our approach for predicting a structure of protein kinase. Nevertheless, we believe that the surface, interior, and secondary structural assignments reported here are sufficiently well developed to provide an interesting and critical test of our approach.

Further, we have no experimental experience with protein kinases, and therefore do not have a biochemist's understanding of this family. Therefore, we have been unable to exploit completely feature (e) of our approach (*vide supra*). This means that we have undoubtedly made some errors in our analysis and overlooked some avenues for refining our structural modeling that would be obvious to those having experience with this system. Nevertheless, we believe that the exercise has been useful. At the very least, it provides a 'worked example' illustrating how we believe that our approach should be applied to predicting protein conformations. Further, should the crystal structure of a member of the protein kinase family not be immediately forthcoming, we will have the opportunity to learn more, and possibly refine the model presented here further.

It is worth noting at this point that several groups have attempted to predict the folded structure of the catalytic domain of protein kinases using other approaches. For example, a prediction by Shoji *et al.* using a Chou-Fasman algorithm found three regions of the catalytic domain with different secondary structures, the first (positions 1–98 in the alignment discussed here) being highly (79%) helical, the second consisting of 3 'subdomains' (positions 99–146, 147–188, and 189–251) each consisting of a beta strand followed by two alpha helices and separated by two beta turns, and the third (252–end) being highly aperiodic (only 18% alpha helix and 20% beta strand) (13). Other predictive work has focused on the fact that the amino terminal portion of the domain has the sequence GXGXXG, a sequence that is conserved in most members of the family. Such a sequence is also found in the 'Rossmann fold', and α-β-α supersecondary structural unit that is present in several proteins that bind nucleotides and dinucleotides. Thus, several authors have suggested that this supersecondary structural unit will be found in the catalytic domains of protein kinases (3).

The hypothesis that protein kinases contain a Rossmann fold was especially attractive to us, as a key element of our strategy for predicting tertiary structure is to find evidence in a secondary structural prediction that a protein belongs to a particular taxonomical class. This assignment, if possible, offers an 'easy' route to further modeling, where the taxonomical class is taken as a working hypothesis to be tested and refined by a variety of methods, including computation and co-variation analysis. Not surprisingly, the extrapolation of a pattern of secondary structure to a particular taxonomical class was central to the assignment of a 'β-barrel' structure to tryptophan synthetase, the first enzyme whose structure was predicted in advance of a crystallographic analysis (14).

Unfortunately, our secondary structural predictions are not consistent with the hypothesis that the protein kinases contain the Rossmann fold as a predominant supersecondary structural motif, nor are they consistent with the classical analysis based on Chou-Fasman, GOR, or other methods mentioned above. Instead, our approach has yielded a secondary structure prediction where consecutive β strands, several aligned to form an *antiparallel* beta sheet, are an important feature. A Rossmann fold, of course, contains a *parallel* beta sheet as its core. Further, the distribution of helices does not follow the pattern suggested by the classical algorithms. These contrasts between predictions made by our method and by other methods are, of course, points where the merits of the various approaches can be critically compared once the 3-dimensional structure of a protein kinase is known.

We have not yet been able to identify with certainty a particular taxonomic class in our secondary structural prediction, although a beta sandwich structure is our preferred choice at the present time. Therefore, an 'easy' route to model building has not been accessible. Instead, we have attempted to assemble the secondary structural elements into a globular structure using distance constraints imposed by our assignments of the active site residues, limited amounts of co-variation analysis that bring distant parts of the polypeptide chain together at specific points, and a variety of data obtained by chemical modification studies.

MATERIALS AND METHODS

The alignment presented by Hanks *et al.* (1) of the sequences of 65 catalytic domains of protein kinase homologs was used as the starting point for this work. Several sequences were added to this alignment to yield an alignment with 79 sequences in all (Fig. 1). These added sequences were checked whenever possible against the original literature or against a computerized data base to remove as many of the inevitable reporting errors as possible. Major insertions in the alignment were excised to save computational time; these segments are invariably assigned as parses (e.g., surface loops) by our approach, and such excision does not affect the overall structural model.

A master alignment is shown in Figure 1, where individual protein sequences are read vertically downwards. The first column of numbers in Figure 1 are the alignment numbers. These are used throughout this manuscript to designate specific positions in the protein sequence. The second column of numbers contains the sequence number from porcine kidney cAMP-dependent protein kinase (15). This is the enzyme for which crystals have been reported, and we presume that this is the numbering that will be relevant to a comparison of the predictions made here with the crystal structure when it emerges. Further, we have divided the proteins in the table into 12 'functional subfamilies', sets of enzymes performing analogous functions. Each column in Figure 1 (including the gaps) has a 'protein number' (from 1 through 90) which is used throughout the text to designate specific proteins.

A key element of our approach is based on the recognition that

	2	3	4	5	6	7	8	9
1	99	95	95	60	62	65	62	56
2		95	95	60	62	65	62	56
3			99	60	62	66	62	57
4				60	62	66	62	57
5					96	79	87	52
6						83	91	54
7							82	53
8								55

	11	12	13	14	15	16	17	18
1	51	52	52	51	50	51	50	50
2	51	52	52	51	50	51	50	51
3	52	53	53	51	50	52	50	51
4	52	53	53	51	50	52	50	51
5	48	49	49	48	47	48	46	48
6	50	51	51	50	49	50	48	50
7	53	54	54	53	52	53	52	52
8	51	53	52	51	50	51	50	52
9	48	48	48	48	46	48	46	48
11		91	92	89	83	99	83	84
12			99	96	83	91	83	82
13				97	83	91	83	82
14					83	89	83	79
15						83	98	78
16							83	84
17								78

	20	21	22	23	24	25	26
1	42	42	43	42	38	41	44
2	42	41	43	42	38	41	44
3	42	41	44	43	39	41	46
4	42	41	44	44	38	41	46
5	38	37	40	40	38	40	42
6	39	38	40	40	39	41	43
7	41	40	41	42	39	41	45
8	40	39	40	40	39	41	44
9	38	38	38	38	37	40	41
11	41	41	38	38	39	40	42
12	40	40	38	37	38	39	42
13	40	40	38	37	38	39	42
14	38	38	36	36	38	38	40
15	38	38	37	37	35	37	40
16	41	41	38	38	38	39	41
17	38	38	36	36	35	36	41
18	42	42	39	38	39	39	42
20		93	48	47	45	44	50
21			48	47	44	44	50
22				94	44	44	46
23					43	43	46
24						62	47
25							46

	28	29	30	31
1	44	44	43	43
2	44	44	42	43
3	44	45	44	44
4	44	44	44	44
5	44	43	41	42
6	44	44	42	43
7	45	42	43	43
8	46	44	42	42
9	42	41	38	38
11	44	41	44	45
12	43	40	42	43
13	43	40	42	43
14	42	39	42	42
15	43	39	42	43
16	44	41	44	44
17	42	39	42	42
18	45	41	42	43
20	50	49	42	42
21	50	48	41	41
22	44	43	42	41
23	43	42	41	40
24	42	42	37	36
25	44	43	41	40
26	46	44	42	42
28		54	52	51
29			48	48
30				89

	33	34	35	36	37
1	33	34	36	31	34
2	33	34	36	31	33
3	33	34	37	31	33
4	33	34	37	31	33
5	33	33	33	31	33
6	33	32	34	32	33
7	33	32	33	33	33
8	34	31	33	34	33
9	32	31	33	30	32
11	30	31	33	30	34
12	29	30	32	30	33
13	29	30	32	30	33
14	28	31	31	29	32
15	29	31	31	30	33
16	30	32	33	30	34
17	29	31	31	29	32
18	31	32	34	32	35
20	36	37	42	37	38
21	37	38	41	37	38
22	34	34	37	37	36
23	35	35	37	36	36
24	34	34	37	32	37
25	34	34	37	33	37
26	34	33	37	33	39
28	40	39	40	38	39
29	35	35	39	36	38
30	34	33	37	36	34
31	34	33	36	36	33
33		69	67	50	47
34			70	52	48
35				51	52
36					44

	39	40	41	42	43	44	45	46
1	32	26	33	38	34	31	34	36
2	32	26	33	37	34	31	34	36
3	32	26	34	38	34	31	34	36
4	32	26	34	37	34	31	34	36
5	31	27	33	35	34	32	35	36
6	31	28	34	36	34	32	36	36
7	32	27	35	38	34	32	34	36
8	32	29	34	36	34	33	35	36
9	31	28	36	34	35	27	32	37
11	33	27	33	35	34	28	34	36
12	33	26	33	35	34	29	34	35
13	33	26	34	35	34	29	34	35
14	33	26	33	34	34	28	32	34
15	31	25	31	34	33	27	31	37
16	33	27	33	35	34	28	34	36
17	31	25	31	34	33	27	31	36
18	32	27	32	36	34	27	33	37
20	35	31	34	34	34	35	37	38
21	34	33	33	34	34	34	38	38
22	34	30	36	37	35	33	35	38
23	35	30	36	37	34	33	34	39
24	34	30	35	36	33	29	33	37
25	37	31	34	38	34	32	34	38
26	35	30	34	37	37	34	36	36
28	35	32	36	35	35	33	41	42
29	35	31	35	35	35	35	43	40
30	35	27	34	37	34	33	38	38
31	34	27	34	36	34	32	37	38
33	38	34	34	33	34	34	33	31
34	36	33	34	36	33	33	34	33
35	38	34	38	35	34	36	37	33
36	33	31	35	32	32	34	36	31
37	36	34	34	34	32	34	35	32
39		33	30	30	32	30	30	32
40			30	27	27	30	30	30
41				47	33	32	37	32
42					34	31	35	34
43						30	33	31
44							31	30
45								37

	48	49	50	51	52	53	54
1	32	33	32	33	30	30	29
2	32	33	32	33	30	30	29
3	33	33	32	33	30	30	30
4	33	33	32	33	30	30	29
5	33	35	35	34	32	30	30
6	34	35	35	35	32	30	31
7	34	36	36	35	29	30	30
8	34	35	35	35	31	30	31
9	34	33	33	33	30	32	31
11	33	33	33	33	29	30	30
12	33	33	33	33	29	30	30
13	33	33	33	33	29	30	30
14	32	32	31	32	28	29	30
15	31	31	31	31	29	29	29
16	33	33	33	33	29	30	30
17	31	31	31	31	29	29	29
18	33	33	33	33	29	30	30
20	32	34	34	34	30	30	30
21	33	34	34	33	30	30	31
22	34	34	34	34	29	28	29
23	33	34	33	33	29	28	28
24	33	34	34	33	32	31	32
25	34	33	33	33	33	32	32
26	34	34	35	34	31	30	29
28	34	35	34	35	32	33	33
29	35	36	35	35	31	31	31
30	34	35	34	34	32	31	32
31	33	33	32	33	33	31	33
33	29	28	28	28	32	30	30
34	29	29	29	28	31	30	30
35	31	32	32	31	31	30	30
36	31	31	31	30	30	29	28
37	31	31	31	30	30	30	30
39	28	29	29	29	28	26	26
40	27	26	26	26	29	28	28
41	36	34	34	34	35	34	35
42	36	36	36	35	33	32	32
43	36	34	34	34	33	33	32
44	31	31	32	31	27	26	26
45	33	32	31	31	27	27	28
46	31	30	30	30	31	33	33
48		83	82	83	35	35	34
49			99	98	36	35	34
50				97	36	35	34
51					35	34	34
52						81	79
53							93

	56	57	58	59	60	61	62	63	64	65
1	35	35	34	36	34	33	34	33	36	34
2	35	35	34	36	34	33	34	33	36	34
3	35	35	34	35	34	34	34	33	36	34
4	35	35	34	36	34	34	34	33	36	34
5	34	35	35	35	34	34	34	34	36	34
6	34	35	35	35	34	34	34	34	36	34
7	35	36	37	36	35	35	35	34	36	34
8	35	35	37	36	34	33	34	33	36	34
9	36	37	38	38	36	35	35	35	37	35
11	36	36	35	37	35	35	36	36	34	36
12	36	35	36	37	36	36	36	36	35	36
13	36	35	36	37	36	36	36	36	35	36
14	35	35	35	36	35	35	35	35	35	34
15	34	34	34	34	34	34	35	35	34	33
16	36	35	35	36	34	35	35	35	34	36
17	34	34	34	34	34	34	34	34	34	33
18	36	36	35	38	35	34	35	36	34	35
20	34	34	34	34	34	34	34	34	33	35
21	34	33	34	34	33	33	33	33	34	34
22	33	32	31	32	31	30	31	31	33	31
23	32	32	30	31	30	30	31	30	32	30
24	39	38	37	38	37	38	38	37	35	36
25	38	38	36	38	35	36	37	36	35	34
26	34	33	34	33	31	30	30	31	33	34
28	37	37	37	37	38	38	38	37	36	35
29	38	38	38	38	38	37	38	37	38	37
30	36	36	36	37	37	37	37	35	36	36
31	35	35	34	36	35	36	36	35	36	34
33	31	31	31	31	30	31	31	30	31	31
34	30	31	30	30	30	32	32	30	31	30
35	34	34	33	33	33	34	35	34	33	33
36	30	29	29	30	27	29	30	28	30	29
37	33	33	33	33	32	32	32	33	30	31
39	29	29	29	29	28	29	29	28	29	27
40	28	27	29	29	27	29	28	27	30	28
41	34	34	32	34	34	33	34	33	36	34
42	36	35	34	36	35	36	37	34	37	34
43	34	36	34	35	35	35	35	34	35	31
44	30	30	30	30	28	29	30	28	29	29
45	33	32	32	33	32	33	34	32	32	31
46	33	33	33	34	34	36	36	34	34	31
48	43	44	44	44	42	41	42	42	44	42
49	42	42	42	43	42	41	41	41	43	40
50	42	43	43	43	42	41	41	41	43	41
51	42	42	42	42	42	40	40	41	43	40
52	38	36	37	37	38	37	38	39	35	34
53	38	38	38	38	39	38	38	40	35	34
54	37	37	37	37	38	37	38	40	35	33
56		91	83	87	73	72	74	75	66	58
57			83	90	74	73	75	77	66	57
58				83	72	69	70	72	65	58
59					75	74	76	76	65	58
60						77	80	86	62	57
61							94	79	60	56
62								82	61	58
63									61	56
64										57

	67	68	69	70	71	72
1	37	38	35	35	34	36
2	37	39	35	35	34	36
3	38	39	35	36	35	37
4	38	39	35	36	36	37
5	36	35	34	37	37	38
6	37	36	34	38	38	39
7	37	37	35	36	35	37
8	37	36	34	37	37	38
9	36	35	35	35	35	35
11	37	36	35	34	34	35
12	38	37	35	36	35	36
13	38	37	35	36	35	36
14	35	35	34	35	34	35
15	33	33	33	34	33	34
16	37	37	34	34	34	35
17	34	34	33	34	33	34
18	37	36	34	36	35	36
20	35	35	35	34	33	34
21	34	34	34	34	34	34
22	32	32	29	33	33	34
23	33	32	30	33	32	34
24	36	36	34	35	35	34
25	34	34	33	35	35	36
26	31	33	31	34	34	35
28	37	35	36	36	36	37
29	37	37	37	36	36	37
30	37	39	37	38	38	40
31	35	37	35	37	37	39
33	32	33	32	32	32	32
34	33	33	33	33	33	34
35	35	37	34	34	35	36
36	30	30	32	32	32	32
37	32	32	32	31	31	31
39	28	30	28	28	28	28
40	26	28	26	29	29	28
41	34	32	31	33	34	33
42	34	34	34	34	34	35
43	32	33	32	35	35	36
44	31	30	30	29	29	29
45	33	33	32	32	32	31
46	34	33	32	34	34	34
48	39	39	40	39	40	40
49	40	40	41	40	40	40
50	40	40	41	40	41	40
51	40	40	41	40	40	40
52	32	33	33	35	35	37
53	33	34	34	35	35	37
54	32	34	34	35	35	36
56	58	60	56	53	53	50
57	59	59	57	53	53	52
58	57	58	57	50	50	49
59	57	58	56	52	52	52
60	58	59	57	50	50	51
61	56	58	57	51	51	50
62	58	59	58	53	52	52
63	57	58	57	50	50	50
64	58	58	57	56	56	55
65	57	59	56	52	53	52
67		81	71	53	53	52
68			72	53	53	52
69				52	52	51
70					98	84
71						83

	74	75	76
1	34	33	35
2	34	33	35
3	34	34	35
4	34	34	35
5	34	32	33
6	33	32	33
7	34	34	33
8	34	34	33
9	33	31	32
11	34	34	32
12	33	33	33
13	33	33	33
14	31	32	32
15	32	32	31
16	34	34	32
17	32	31	31
18	35	34	33
20	32	32	33
21	30	31	33
22	31	31	31
23	31	30	30
24	31	32	32
25	31	32	31
26	31	31	33
28	34	33	33
29	35	34	36
30	35	34	34
31	34	33	32
33	32	31	29
34	31	31	29
35	32	31	32
36	28	29	27
37	31	31	30
39	26	27	25
40	26	26	27
41	30	31	30
42	35	35	31
43	30	30	29
44	29	28	27
45	31	32	32
46	32	32	31
48	37	38	36
49	38	38	36
50	38	38	36
51	38	38	36
52	30	30	32
53	30	30	32
54	30	30	32
56	46	49	46
57	46	48	46
58	45	48	46
59	46	49	45
60	45	46	46
61	45	46	45
62	45	47	45
63	45	46	45
64	45	49	48
65	45	48	46
67	50	50	49
68	49	50	50
69	48	49	48
70	43	45	42
71	43	45	42
72	43	44	43
74		85	64
75			65

	78	79	80	81	82	83	84	85
1	31	31	30	30	31	30	32	32
2	32	32	31	30	31	31	32	32
3	32	33	31	30	31	31	33	31
4	33	33	31	30	31	31	33	31
5	33	30	32	32	31	33	33	34
6	34	31	32	32	31	33	33	33
7	32	31	31	31	30	32	34	34
8	33	31	32	32	31	33	34	35
9	32	31	31	32	30	33	32	33
11	32	31	30	29	30	30	34	32
12	33	31	31	29	30	30	35	33
13	33	31	31	29	30	30	35	33
14	32	31	30	28	28	28	34	32
15	32	32	30	27	28	28	33	30
16	32	31	30	29	30	30	34	32
17	32	31	30	27	28	28	33	30
18	34	33	32	29	30	31	36	32
20	31	31	30	31	31	33	31	33
21	30	30	29	30	30	32	31	33
22	29	28	30	28	29	27	30	30
23	30	28	30	28	28	28	30	30
24	30	30	31	33	33	33	31	33
25	31	30	30	31	32	32	32	32
26	32	30	31	31	32	32	31	30
28	34	34	34	31	31	31	36	34
29	35	34	33	33	33	33	35	34
30	35	35	34	31	32	32	39	36
31	34	34	33	31	31	31	37	34
33	31	30	31	29	30	30	32	31
34	28	29	28	28	26	29	29	30
35	31	33	31	30	29	32	31	32
36	30	28	27	27	26	29	30	29
37	30	29	29	28	28	26	29	32
39	28	26	25	24	24	25	26	27
40	25	24	24	25	23	25	27	28
41	29	29	29	28	27	29	31	30
42	31	31	30	32	31	31	32	31
43	32	31	32	31	31	31	33	30
44	27	26	26	27	27	26	30	29
45	29	29	27	29	29	29	30	33
46	30	30	32	29	29	31	32	32
48	37	37	35	36	35	34	37	38
49	36	36	36	36	36	35	37	38
50	36	36	37	36	36	35	37	38
51	35	36	36	36	36	35	37	37
52	33	34	33	31	31	34	32	32
53	31	33	31	34	31	33	32	32
54	31	33	31	34	31	33	31	32
56	48	47	46	50	50	48	46	46
57	47	47	45	51	50	48	45	45
58	47	46	43	48	47	47	46	46
59	46	46	44	50	48	47	46	46
60	47	47	44	50	48	46	48	48
61	46	46	43	49	47	46	47	45
62	46	46	43	50	48	47	47	46
63	47	46	44	50	49	45	47	44
64	46	45	45	47	48	47	46	48
65	45	44	44	49	48	48	52	47
67	50	51	49	49	49	47	48	50
68	47	48	46	50	49	47	52	49
69	49	50	49	49	49	49	48	48
70	50	50	48	51	50	48	50	48
71	50	50	48	51	50	48	50	48
72	50	49	48	49	48	49	50	49
74	44	44	43	45	46	47	44	49
75	45	44	44	45	47	47	45	50
76	41	42	41	45	46	45	42	47
78		84	70	56	57	54	53	49
79			71	56	56	56	52	49
80				52	54	51	51	49
81					78	61	50	48
82						62	49	47
83							49	46
84								46

	87	88	89	90
1	33	31	33	34
2	33	31	33	34
3	32	32	34	34
4	32	31	33	33
5	34	34	35	34
6	34	34	35	33
7	33	34	34	34
8	33	34	34	33
9	34	36	36	32
11	33	34	35	33
12	34	34	35	33
13	34	34	35	33
14	34	34	35	32
15	31	31	33	33
16	33	34	34	33
17	31	31	33	32
18	33	34	35	34
20	32	31	32	34
21	31	31	31	32
22	29	29	30	30
23	29	29	30	30
24	34	34	34	34
25	32	32	32	32
26	31	31	33	32
28	34	35	35	37
29	35	36	36	34
30	36	35	37	36
31	34	34	36	34
33	29	30	30	32
34	30	30	31	30
35	30	29	31	33
36	29	28	29	30
37	29	30	31	31
39	27	26	27	27
40	26	27	26	29
41	35	32	33	34
42	33	33	34	36
43	35	36	36	34
44	27	26	27	26
45	32	32	32	30
46	31	32	32	33
48	40	38	39	38
49	39	38	40	38
50	39	39	40	38
51	38	38	39	38
52	34	32	32	32
53	35	33	32	32
54	34	32	31	33
56	47	47	46	47
57	46	48	46	47
58	47	47	47	45
59	47	48	47	46
60	50	49	48	47
61	49	48	47	50
62	48	48	46	49
63	49	50	50	48
64	49	48	48	48
65	48	47	45	46
67	47	48	48	50
68	48	47	48	49
69	47	48	48	49
70	46	49	47	48
71	47	49	47	48
72	46	46	46	48
74	44	44	44	48
75	46	45	45	48
76	44	44	44	43
78	47	49	50	51
79	46	47	46	52
80	44	46	45	51
81	48	50	47	52
82	47	51	48	50
83	47	47	44	47
84	50	51	50	50
85	48	46	48	47
87		69	67	54
88			76	56
89				57

FIG. 2. Matrix showing percentage pairwise identities between individual protein kinase sequences in Figure 1.

information regarding folding patterns appears and disappears at different points during the divergent evolution of a set of proteins. Thus, the 'minimum pairwise identity' (or MPI) within a subgroup of proteins must be known to evaluate the significance of any conservation or divergence seen within the subgroup. A matrix showing the pairwise identity of each sequence used here is shown in Figure 2. The MPI values for the individual functional subfamilies are given in the caption to Figure 1.

RESULTS AND DISCUSSION

A central feature of our approach is the use of a set of algorithms with different accuracies to assign positions to the surface or the interior of a folded protein. In assigning secondary structure, the different outputs of the different algorithms are then used in order, starting with the most reliable assignments and proceeding towards the least reliable. Thus, the secondary structural prediction is made from the most reliable data if possible; only if the most reliable data are inadequate to yield an unambiguous assignment of secondary structure are the less reliable data considered.

Figure 3 shows, by alignment number, the surface, interior, and parsing assignments. Together with the alignment of sequences in Figure 1, Figure 3 contains the central body of data on which further analysis is based. Surface positions are identified by the presence of variation in more than one subgroup at different levels of minimum pairwise identities

			Number of Variable Subgroups						
	≥10	9	8	7	6	5	4	3	2
001	11.5.8	4.8			7.5	9.4;8.4			
002			5.7	6.6;4.6					9.2;8.2;7.2
003	INSIDE								
004	11.6.6		4.6	7.4			8.3		
005				4.2		5.2			
006	13.6.4	4.4	8.3						
007	10.5.9	4.8	6.7				7.3		
008					5.5;4.6		8.4;7.4;6.4		
009	INSIDE								
010	*secondary parse*								
011	10.4.5								
012	*secondary parse*								
013						4.2			
014	INSIDE								
015									
016			4.6			5.4		6.3	7.2
017	INSIDE								
018					4.2				
019					4.6			5.2	
020	INSIDE								
021	10.5.6			6.4;4.5	8.3;7.3			9.2	
022	inside						4.2		
023	10.6.7		4.6	7.6		8.4		9.2	
024		5.7		6.5;4.6	7.4	8.3			
025	*PRIMARY PARSE*							5.2;4.3	
026	*PRIMARY PARSE*							4.2	
027	*PRIMARY PARSE*						6.2;5.2	4.2	
028	*PRIMARY PARSE*						6.4;5.4	4.3	8.2;7.2
029	*PRIMARY PARSE*							5.2	4.2;6.2
030	*PRIMARY PARSE*								
031	*PRIMARY PARSE*								
032	*PRIMARY PARSE*								
033	*PRIMARY PARSE*								
034	*PRIMARY PARSE*								
035	*PRIMARY PARSE*								
036	*PRIMARY PARSE*								
037	*PRIMARY PARSE*								
038	*PRIMARY PARSE*								
039				5.2	4.4				
040		5.7	6	7.4;4.5	8.4		9.3		
041	10.5.6		6.5	4.6			7.3	9.2;8.2	
042		6.4;5.5		4.5					
043	INSIDE								
044	INSIDE								
045	INSIDE								
046	Active Site								
047			4.3						
048	INSIDE								
049						4.5	5.4		
050			4.6		5.4				
051			4.6		5.3	6.3	8.5;7.2		
052			5.3;4.3	6.2					
053		4.6	5.2			6.2			
054			4.5	5.2		6.3			
055	10.4.7		5.5		6.4	7.3		8.2	
056	*PRIMARY PARSE*							4.2	
057	*PRIMARY PARSE*								
058	*PRIMARY PARSE*								
059	*PRIMARY PARSE*							4.2	
060				4.5		5.2			
061			5.4;4.6			6.2			
062				5.5;4.5	6.4	7.4			
063				4.6		5.5	7.3;6.4	8.2	9.2
064	INSIDE								
065		4.3			5.3				
066			5.7		4.6		7.3;6.3		8.2
067	Active Site								
068	INSIDE								
069			4.6		5.4		6.3	7.2	8.2
070	INSIDE								
071	INSIDE								
072					4.6				7.2;6.2;5.2
073			5.4;4.5		6.3		7.2		
074	INSIDE								
075	*PRIMARY PARSE*						4.3		6.2;5.2
076	*PRIMARY PARSE*								6.2;5.2;4.2
077	*PRIMARY PARSE*								
078	*PRIMARY PARSE*								
079	*PRIMARY PARSE*								
080	*PRIMARY PARSE*								
081	12.6.12	4.9	8.7;7.8					9.3	
082						5.2;4.2	6.2		
083	11.5.6		6.4;4.7						
084					4.6		5.4	6.2	
085	INSIDE								
086	INSIDE								
087				5.6;4.6		6.2	8.2;7.2		
088	INSIDE								
089	INSIDE								
090	inside						4.2		
091	INSIDE								
092	INSIDE								
093	INSIDE								
094	10.5.7			6.5			7.3	8.2	
095				4.7	5.5		6.2		
096	*PRIMARY PARSE*							4.2	
097	*PRIMARY PARSE*								
098	*PRIMARY PARSE*							4.3	
099	*PRIMARY PARSE*								
100	*PRIMARY PARSE*								
101				4.5		5.2	6.2		7.2
102				4.5	5.4		6.3	7.2	
103	INSIDE								
104	INSIDE								
105	inside								9.2
106	INSIDE								
107	INSIDE								
108									4.2
109	INSIDE								
110	INSIDE								
111			6.6;5.7		7.3;4.6		8.2		
112			5.4	6.4	7.3;4.3		8.2		
113	*PRIMARY PARSE*								
114	*PRIMARY PARSE*								
115							4.4		6.2;5.2
116	INSIDE								
117						4.3			
118				6.4;5.4	4.6				
119	INSIDE								
120	INSIDE								
121					4.5		5.3		6.2
122		5.7	4.8	6.5			7.3	8.2	9.2

FIG. 3. Surface, parsing, inside and active site assignments for protein kinase.

FIG. 3. *Cont'd*

123			5.3;4.5		6.3		7.2				8.2	
124		4.7	6.6;5.6			7.5				8.3	9.2	
125	10.5.8	6.8;4.8					7.4		8.3		9.2	
126	11.5.7			7.5				8.3				
127						6.3;5.3	4.2			7.2		
128	*PRIMARY PARSE*											
129	*PRIMARY PARSE*											
130	12.5.7	7.6;4.7;6.7			8.4				9.3			
131	**INSIDE**											
132		5.5;4.7			6.3							
133	11.5.6		6.4;4.5		7.4	8.4						
134	**INSIDE**											
135	10.5.2	6.2			4.2							
136			5.5		4.6		6.2					
137	**INSIDE**											
138	**INSIDE**											
139					4.4							
140									4.3		5.2	
141	**INSIDE**											
142	**INSIDE**											
143					4.7	5.5			6.3			
144												
145	**INSIDE**											
146		4.8	5.6				6.4			7.2		
147	**INSIDE**											
148	**INSIDE**											
149										4.2		
150					4.6	5.5	6.4		7.3		8.2	
151			5.7		7.7;6.7;4.6		8.5				9.2	
152					4.7	5.5			6.3			
153	**INSIDE**											
154	**INSIDE**											
155	**INSIDE**											
156												
157	Active Site											
158	**INSIDE**											
159												
160	*secondary parse*											
161									4.3			
162	Active Site											
163	**INSIDE**											
164	**INSIDE**											
165	**INSIDE**											
166					4.5	5.2						
167	11.6.8	7.6	4.6			8.4						
168	10.5.9		4.7			6.6			7.4	8.3	9.2	
169					5.5;4.5	6.4				7.2		
170	*PRIMARY PARSE*											
171	*PRIMARY PARSE*											
172	*PRIMARY PARSE*											
173	*PRIMARY PARSE*											
174	*PRIMARY PARSE*											
175	*PRIMARY PARSE*											
176	*PRIMARY PARSE*											
177			4.3									
178	**INSIDE**											5.2;4.2
179	**Inside**											
180	**INSIDE**											
181	**INSIDE**											
182	Active Site											
183	**INSIDE**											
184	*secondary parse*											
185	**INSIDE**											
186	**INSIDE**											
187									4.4		7.2;6.2;5.2	
188							6.2;5.2		4.3		7.2	
189					4.2							
190	12.5.10	4.6					8.4			9.2		
191		4.5	5.3		6.3		7.2					
192					5.7;4.6				7.4;6.4		8.2	
193		4.7				5.4			6.2	7.2		
194						4.5			5.3	7.2;6.2		
195	*PRIMARY PARSE*						4.3		6.3;5.3	7.2		
196	*PRIMARY PARSE*											
197					4.4	5.2						
198					4.4	6.3;5.3			7.2	8.2		
199						4.3						
200			4.4			5.2						
201					4.3				5.2			
202	**INSIDE**											
203	*secondary parse*											
204	*secondary parse*											
205										4.2		
206	**INSIDE**											
207	**INSIDE**											
208	**INSIDE**											
209	*secondary parse*											
210	Active Site											
211	**INSIDE**											
212	**INSIDE**											

FIG. 3. *Cont'd*

213	*PRIMARY PARSE*								4.2
214	*PRIMARY PARSE*								
215	*PRIMARY PARSE*								
216	*PRIMARY PARSE*								
217				5.4	6.3;4.5				
218			4.6	5.3	6.3			7.2	
219			4.7			5.4	6.3		9.2;8.2;7.2
220	10.5.4	4.4	6.4						
221	**INSIDE**								
222									9.2
223				4.3	5.2				
224	*PRIMARY PARSE*								
225			4.7	5.3					
226	**inside**								
227	CM1D								
228	**INSIDE**								
229	**INSIDE**								
230	**INSIDE**								
231	**INSIDE**								
232	*secondary parse*								
233	**INSIDE**								
234	**INSIDE**								
235	**INSIDE**								
236	**INSIDE**								
237									
238	**INSIDE**								
239	**INSIDE**								
240	**INSIDE**								
241	*PRIMARY PARSE*								4.2
242							4.3		6.2;5.2
243		4.7	6.5;5.5				7.3		
244			5.2	4.2					
245	*secondary parse*								
246	**INSIDE**								
247				5.2;4.2					
248					4.5	5.3	6.2		
249	*PRIMARY PARSE*								
250	*PRIMARY PARSE*								
251	*PRIMARY PARSE*								
252	*PRIMARY PARSE*								
253	*PRIMARY PARSE*								
254	*PRIMARY PARSE*								
255	*PRIMARY PARSE*								
256	*PRIMARY PARSE*								
257	*PRIMARY PARSE*								
258	*PRIMARY PARSE*								
259	*PRIMARY PARSE*								
260					6.2;5.4;4.5			7.2	
261			5.7	4.6	6.5		7.3		
262				4.5	6.3;5.3		7.2		
263			5.5	4.6	6.4		7.2		
264	10.5.8	4.7	6.6			7.3	8.2		
265	**INSIDE**								
266	**INSIDE**								
267		5.8	6.7;4.8		7.5		8.4		
268	10.6.3			4.2		7.2			
269	**INSIDE**								
270				5.3		6.3;4.3		7.2	
271	11.5.11		7.8;4.7			8.5			9.2
272									
273				4.4	5.2				
274	10.5.8	4.8	6.6		7.5				
275						4.2			
276	12.5.7	6.5	4.7			7.4	8.3		
277			6.3;5.3		4.3				
278	*secondary parse*								
279	11.5.8	4.7	6.6						
280				5.2	4.5				
281	*PRIMARY PARSE*					4.2			
282	*PRIMARY PARSE*								
283	*PRIMARY PARSE*							4.3	5.2
284	*PRIMARY PARSE*					4.4		5.2	
285	*PRIMARY PARSE*							4.2	
286	**INSIDE**								
287					4.2	5.2		6.2	
288	11.5.10	6.9	7.8;4.7			8.5			
289		5.9	4.8	6.7			7.4		8.2
290	**INSIDE**								
291		6.4;5.4		7.3;4.3					
292	10.5.9	6.9	7.7	4.6	8.4		9.2		
293	**INSIDE**								
294	**INSIDE**								
295	11.5.7	6.7;4.6		8.5;7.5			9.2		
296	12.5.11	4.9	7.6		8.5			9.2	
297	**INSIDE**								
298	**INSIDE**								
299	11.5.7	4.6		6.5				7.3	8.2
300			5.4;4.5		7.4;6.4				
301	10.5.8	6.7;4.8		7.5				8.3	
302			4.2						

FIG. 3. *Cont'd*

303	14.5.9			8.5		9.2			
304	10.5.7	6.7	4.6				8.3;7.3		
305	<u>Active site</u>								
306	*secondary parse*								
307			5.3;4.3		6.2				
308	**INSIDE**								
309	10.5.7	4.7	6.5			7.3			
310	10.5.7		4.8	6.5			7.2		
311					7.3;6.3;5.2				
312	**INSIDE**								
313	11.5.4			7.2		8.2			
314	10.6.6				7.4;4.5			8.2	
315	*PRIMARY PARSE*								
316	*PRIMARY PARSE*								
317	*PRIMARY PARSE*								
318	*PRIMARY PARSE*						7.4;6.4	9.3;8.3;5.3	4.2
319			5.4	4.5	6.4			7.2	
320	12.5.8	4.7		7.2			8.2		
321	**INSIDE**								
322		5.2		4.3					
323	11.5.7				7.4		8.3		
324	11.5.9	6.7		4.7	7.4	9.3	8.2		
325			5.3		6.2;4.5				

FIG. 3. *Cont'd*

FIG. 3. Assignment of positions by alignment number in the folded catalytic domain of protein kinase. The figure is ordered to indicate (a) by the column in which an entry appears, how many variable subgroups can be found at a particular position; (b) by the first number in the entry, the number of the cluster in which that number of variable subgroups is found (cluster 9 contains subgroups with MPI = 90%, cluster 8 contains subgroups with MPI = 85%, cluster 7 contains subgroups with MPI = 80%, cluster 6 contains subgroups with MPI = 70%, cluster 5 contains subgroups with MPI = 60%, cluster 4 contains subgroups with MPI = 50%); and (c), by the number following the point, the number of variable subgroups containing at least one polar residue (D, E, K, N, R). The greater the number of variable subgroups (i.e., the farther to the left in the figure an entry appears), the higher the cluster number, and the larger the number of subgroups containing a polar residue, the stronger the surface assignment. When there are more than 9 variable subgroups at the indicated position, the figure lists first the number of variable subgroups in the strongest surface assignment. Where no surface assignment is made, and where other criteria hold (4,5) the position is strongly assigned to the inside, and indicated with the bold label INSIDE. Weaker inside assignments are indicated in lower case letters (e.g., inside). Parsing elements are indicated in italics.

(see References 4 and 5). The column in which an entry is made in Figure 3 indicates the number of variable subgroups in a particular cluster of subgroups with a particular MPI (4, 5), with the MPI of the subgroups designated by the first number in the entry, followed by a period and a number indicating the number of these subgroups that contain at least one polar residue (D, E, K, N, R). Due to the large number of aligned sequences, the assignments should be quite accurate (>95%) and relatively complete (>95% of all surface residues identified). These predictions can also be compared with the crystal structure when it becomes available.

These data are used to construct secondary structural hypotheses for parsed segments of the alignment. The parsing of the alignment is discussed below, as are the detailed considerations leading to a secondary structural assignment. Nevertheless, it is immediately evident that several of the segments almost certainly fold to give surface helices. Figure 4 shows Edmundson helical wheels of the regions of the protein that are assigned as

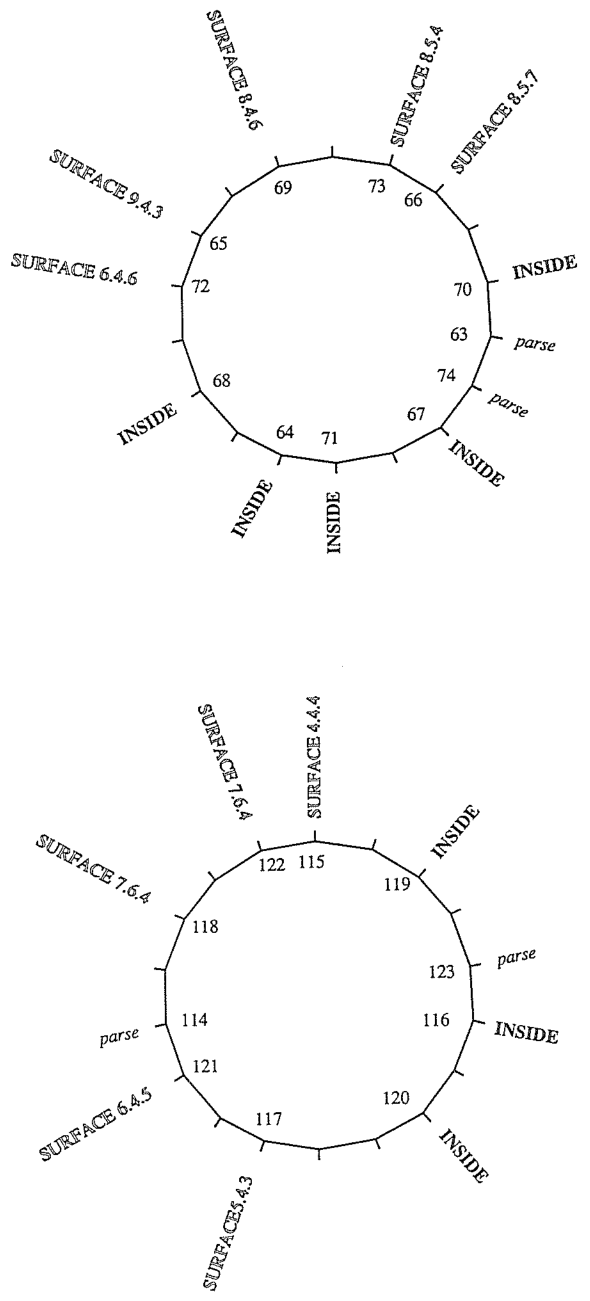
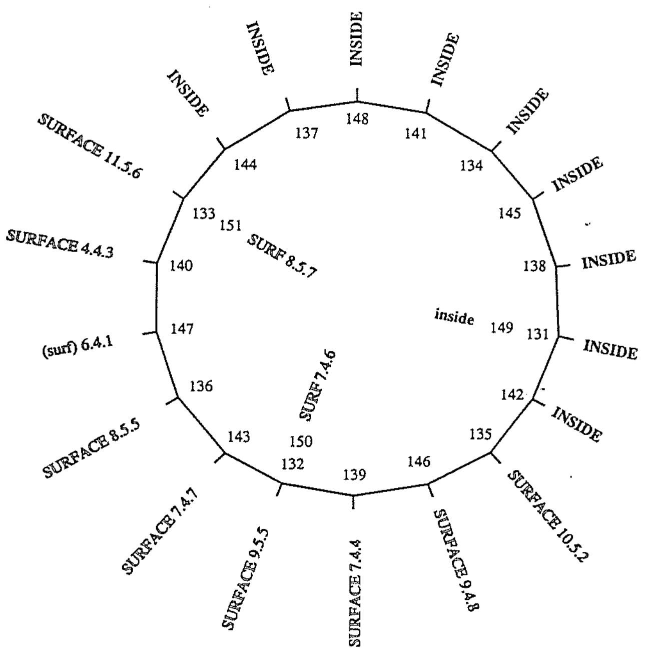


FIG. 4.



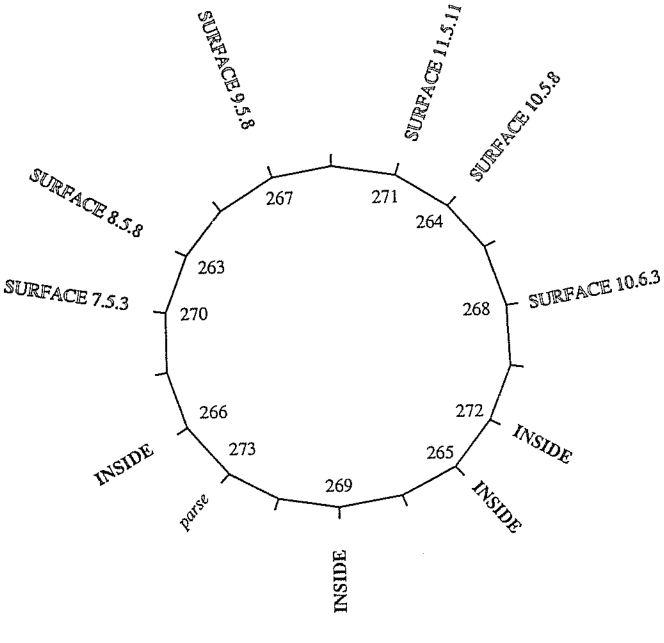


FIG. 4. *Cont'd*

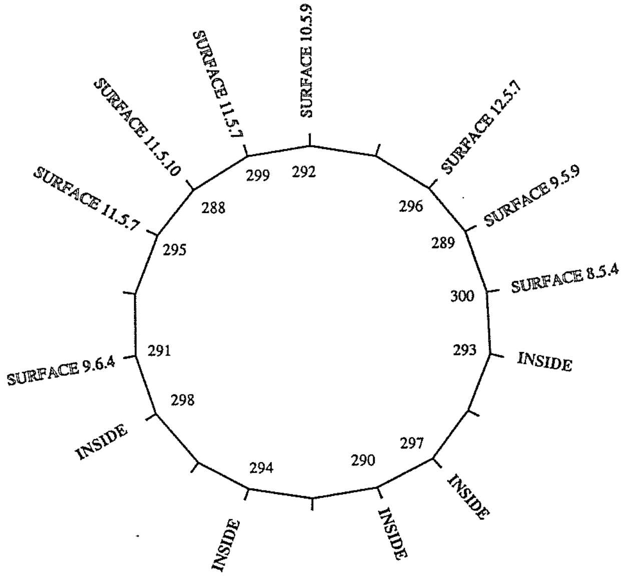


FIG. 4. *Cont'd*

FIG. 4. Edmundson helical wheels of the segments of the protein kinase alignment that are assigned helical conformations. Strong surface assignments are written in upper case letters, weaker assignments in lower case letters or (if very weak) in parentheses.

surface helices. One helix is remarkably long, considering the expected size of the globular structure (ca. 26 Å) (16). Helical wheels are designated by inside residues (boldface) and surface residues (outline, followed by x.y.z, where x is the number of variable subgroups in cluster y, with z the number of subgroups with a polar residue (D, E, K, N, R)).

A complete secondary structural assignment is presented in Figure 5. Detailed discussions for each segment are given below.

Assembly of the secondary structural elements to form supersecondary structures requires that distant positions in the polypeptide chain be brought together in three-dimensional space. One way of obtaining information to assemble these elements comes from assignments of certain positions in the alignment to the active site, as these must be brought together in the folded protein. Unfortunately, the active site of protein kinase is quite large; therefore, these distance constraints are not very demanding. Nevertheless, active site assignments at positions 46, 113, 116, 156–162, 182, 199, 201, 210, 237, and 305 have been used to assemble proposals for the folded structure shown in Figure 6.

Further distance constraints can be found using 'co-variation analysis' that

unit	maximum positions	minimum positions	positions cAMP-PK	secondary struct.	max length	min length	cAMP PK	comments
1	001-012	001-011	001-011	coil	12	11	11	not active site (?, chem. modification)
2	012-023	014-021	012-022	beta	12	8	11	bend positions 15-16
3	023-042	025-041	023-041	coil	20	0	5	not active site
4	041-049	042-048	042-048	beta	9	7	7	active site, bend 46
5	049-063	050-059	049-060	coil	15	1	10	not active site
6	060-075	064-073	061-074	alpha	16	10	14	active site 67
7	074-083	075-083	075-083	coil	10	3	3	not active site
8	083-095	084-092	084-093	beta	13	9	10	bend position 87
9	093-103	096-100	094-102	coil	11	0	8	not active site
10	101-111	103-107	103-111	beta	11	5	9	bend position 108
11	108-115	112-114	112-114	coil	8	1	2	active site 113 & 116
12	115-125	115-122	115-124	alpha	11	8	11	confirmed in subfamilies
13	123-132	126-128	126-130	coil	10	0	2	not active site
14	129-160	133-151	131-152	alpha	32	19	22	surface helix
15	152-160	omitted	153-156	beta	9	0	4	weak assignment
16	152-162	161-162	157-162	active site	11	2	6	active site coil
17	163-168	163-165	163-166	beta	6	3	4	
18	166-177	169-176	167-176	coil	12	0	3	not active site
19	177-187	178-184	177-186	beta	11	8	10	active site bend 182
20	185-191	omitted	omitted	alpha	7	0	0	divergent conformation
21	185-204	192-197	187-200	coil	20	2	10	active site coil
22	198-217	205-208	201-212	beta	20	4	12	bend 203-204 & 208-209, act. site 210
23	209-225	218-224	213-225	coil	16	5	8	not active site
24	225-243	226-240	226-241	beta	19	15	16	bends positions 232 237
25	241-261	244-259	242-261	coil	20	4	8	not active site
26	260-275	263-272	262-274	alpha	16	10	13	realign alignment
27	273-288	276-284	275-285	coil	16	0	6	not active site
28	284-301	288-300	286-301	alpha	18	13	16	
29	301-313	302-306	302-306	coil	13	5	5	active site (?) coil
30	307-313	omitted	307-312	beta	7	0	6	weak assignment
31	313-318	314-318	313-318	coil	6	1	6	not active site
32	319-325+	319-325	319-325	coil	7+	7	7+	possbile short helix

FIG. 5. Secondary structure prediction, by alignment number, for the protein kinase family. The maximum length of each segment, the minimum length, and the preferred length are noted. The preferred length is presented for the cAMP functional subfamily, as this is the subfamily that is presently the subject of crystallographic studies.

detects positions far apart in the sequence where the amino acids at these positions co-vary.

Finally, these purely theoretical predictions can be combined with experimental work that uses chemical modification to identify regions of the protein that are involved in binding ATP and protein substrates. This experimental information is often redundant given information concerning the positions of active site residues obtained using our method. However, as certain assignments of active site residues (in particular, assignments to the active site of highly conserved charged residues) are not highly reliable, this extra information is valuable to confirm active site assignments made by our approach.

Together, several slightly different proposals for the folded structure of the protein can be derived. These are shown in Figure 6. Unfortunately, protein kinases lack sulfide bonds that might be used to provide distance constraints connecting positions of the polypeptide chain that do not lie at the active site. Finally, we have no general knowledge of the biochemistry of protein kinases, and therefore cannot use special knowledge to constrain distances in this protein, as was done, for example, in alcohol

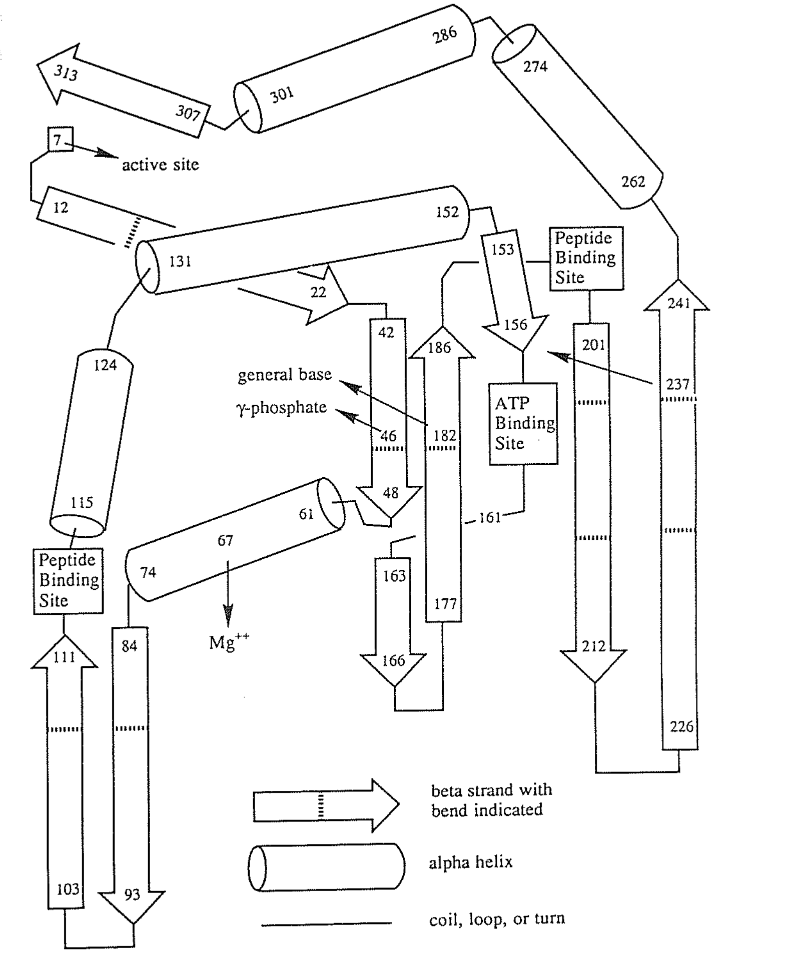


FIG. 6. Folded structures consistent with the secondary structure prediction presented in Figure 5, and constraints imposed by assignments of positions to the active site, co-variation analysis, and experimental data.

dehydrogenases (see Reference 4), a protein with which we are extremely familiar. Thus, the structural models shown in Figure 6 are not complete.

In this modeling effort, we have not taken advantage of any of the many good methods for analyzing structure that are already known in the literature (7). It is conceivable that pattern recognition, statistical analyses, or direct computation might all improve on the structural model that we have proposed, and we would like to encourage those involved in developing these methods to use our structural model as a working hypothesis for applying these methods. Further, we have found evidence that in segments of the alignment, the folded structure of different protein kinases is different (most seriously between positions 190 and 225). This implies that refinement of structure in these regions *must* involve methods other than the one used here.

Is this structural model correct? We cannot say at this time. Circular dichroism spectra have suggested that the cAMP-dependent protein kinase from rat contains 49% alpha helix and 20% beta sheet when not containing a peptide ligand, and 31% alpha helix and 55% beta sheet when bound to a peptide substrate (35). The prediction that is made by our method suggests that 32% of the sequence is alpha helix and 34% is beta strand. The agreement is not unsatisfactory, but cannot be taken as any but the weakest of indicators. Evaluations of folded structure using circular dichroism spectra are quite imprecise. Random assignment of secondary structures would give similar values to those obtained by CD, and, of course, the test of the model is whether it correctly predicts *which* segments adopt which conformations. Therefore, this discussion can only be concluded following the emergence of a crystal structure.

Detailed Discussion of Secondary Structure

A glossary of abbreviations and terms that are necessary for understanding the discussion that follows are given below.

Polar residue: In this discussion, Asp, Glu, Arg, Lys, and Asn.

Hydrophobic residue: In this discussion, Phe, Ile, Leu, Met, Val, Trp, and Tyr.

APC: All positions conserved, indicating that all proteins have the same amino acid at this position.

CMX: Count minus X, a position where all but X proteins have the same amino acid.

MPI: Minimum pairwise identity. The significance of conservation or variation within a subset of proteins depends on their overall sequence identity. Conservation is, of course, most significant in proteins whose sequences are highly divergent; divergence is most significant between proteins whose sequences are otherwise highly conserved. The MPI value for a subgroup of proteins in the alignment is the percent sequence identity of the two least similar proteins in the subgroup.

Clusters of subgroups with an MPI = XX%: The proteins in the alignment are placed on an evolutionary tree, where they are divided into subgroups with different overall levels of sequence identity. All subgroups with a particular MPI belong to a particular cluster of subgroups. In this discussion, cluster 9 contains subgroups with MPI = 90%, cluster 8 contains subgroups with MPI = 85%, cluster 7 contains subgroups with MPI = 80%, cluster 6 contains subgroups with MPI = 70%, cluster 5 contains subgroups with MPI = 60%, cluster 4 contains subgroups with MPI = 50%, cluster 3 contains subgroups with MPI = 30%, cluster 2 contains subgroups with MPI = 30%, and cluster 1 is reserved for the entire alignment.

Functional subfamilies: The proteins in the alignment are divided into 12 functional subfamilies, where members of each functional subfamily are distinct from members of other functional subfamilies by virtue of the type of regulatory role they perform or the substrate that they act upon.

Parse: A segment that divides the alignment into segments whose secondary structure is considered separately.

Parsing string: A sequence of consecutive amino acids in a protein that indicates that the segment lies between standard secondary structural units.

Distributed parse: A parse built from parsing elements that appear in different subgroups of the alignment at neighboring position numbers.

Splits: Positions where the amino acid is conserved within subgroups, but different between subgroups, where the pattern of variation is 'tree-like'.

String: A set of consecutive positions in the alignment.

Non-standard secondary structure: Other than an alpha helix or a beta strand.

Hydrophobic anchor for an external loop: A position with hydrophobic amino acids in most proteins appearing in a segment that is a parse or otherwise assigned as a surface loop.

Reflexivity: A position displays reflexivity when the pattern of variation involving particular amino acids is the same in two distant subgroups, so that the variation is not 'tree-like'.

Functional variable: A position with more than one variable subgroup, where at least some amino acids in the subgroup bear functional groups (C, H, Q, S, T) but no polar groups (D, E, K, N, R).

Hydrophobic variable: A position with more than one variable subgroup, but where none of the amino acids in the subgroup bear functional groups (C, H, Q, S, T, D, E, K, N, R).

This analysis was completed and submitted for publication on September 21, 1990, and the substance of the prediction in this manuscript has not been altered since. At this time, we had no crystallographic information regarding the structure of protein kinases.

Parsing elements. The core of any structural analysis is a discussion that, from the beginning to the end of an alignment, examines every structural feature of the protein family, one position at a time. This discussion is, of course, tedious, but any attempt to abbreviate it is chemically naive. To simplify the discussion, we divide the alignment into manageable pieces using 'parsing elements', or 'parses'. Parsing algorithms are intended to identify segments that lie between standard secondary structural units (e.g., α helices or β strands), allowing the biological chemist to consider the secondary structure of the segments between parses individually.

Parsing algorithms have been developed with different degrees of reliability. In this discussion, the large size of the alignment allows us to use only the strongest parsing algorithms first, the 'primary parses'. The strongest of the primary parses are regions where the alignment shows that one or more proteins in the family has undergone an insertion or deletion. During divergent evolution, insertions (or, conversely, deletions) are generally not made within standard secondary structural elements. With an alignment as large as ours, primary parses alone are almost entirely adequate to divide the alignment into manageable segments, although, it will be seen (*vide infra*), several of the parsed segments are rather large considering the presumed dimensions of the protein. In these cases, it is appropriate to search for secondary parses within the parsed segment. Secondary parses are discussed when the parsed segments are discussed (below).

The larger the alignment, the more likely that a deletion will have arisen within a secondary structural element (or, with the same effect, a misalignment will have incorrectly placed a deletion in a secondary structural element). These will create errors in parsing assignments. Further, approximately once every 400 positions in a typical alignment, a deletion will occur within a secondary structural element (for example, in alcohol dehydrogenase (ADH), a deletion at alignment position 268 (position 256 in the horse liver enzyme) is in a helical region). Finally, the alignment is the weakest in regions of surface loops (which are normally the most rapidly diverging segments of a protein). Thus, assignment of parses is not automatic; it requires answers to three issues:

(a) Where are the *beginning* and *end* of the parse?

(b) Is the parse *confirmed* by segments within the insertion?

(c) Can the deletion or insertion be removed by *shifting* of the alignmen itself?

To elaborate briefly on these points, inserted segments that adop a non-standard secondary structure are often characterized by unusua 'parsing strings', consecutive amino acids that indicate a non-standar structure. 'Non-standard' denotes a structure that is neither an α helix no a β strand; such structures need not, however, be 'random'. Consecutive G PXP motifs, and PG strings are common in external loops that are pron to deletion during divergent evolution, although more exotic strings ar known. For example, in the alignment of receptors families that include the muscarinic receptor, such strings include SSSSS, PPPPALPPPP PXPXPXPXP, and NNNN. A parsing element is said to be *confirmed* if inserted segments contain strings of this sort.

Point (c) addresses errors in the alignment itself. Again, virtually all alignments contain errors, the juxtaposition of amino acids in different proteins that are not true homologs, that is, are not descendent from the same codon in a common ancestor. These errors create significant problems at all levels of our analysis, and the alignment therefore must be reevaluated throughout the structural modeling. Examination of the alignment focuses on *alignment anchors*, positions where the conservation across the entire alignment is sufficiently high that there can be no question that the alignment is correct at this position. The alignment is then built out from the anchors towards other anchors, as exemplified below.

Rather than discussing these points further in the abstract, we examine the primary parses in the kinase alignment in detail. The parses are designated by position number, with the bold face numbers inside the minimum possible extent of the parse, and the outside number (plain type) the greatest possible extension of the parse.

*Primary Parse 1: 23–**25–41**–42*. The first deletion parse is associated with a major insertion at positions 25–38. The parse is confirmed by parsing strings in some of the inserted sequences. For example, in protein 80, there is a PPNG sequence. The parse is reinforced at its amino end by P scattered in positions 23 and 24 in proteins where there is no insertion. It is reinforced at its carboxyl end by scattered P and deletions at positions 39–42. Further, alignment anchors at positions 44 (CMI A) and 46 (APC K) make the alignment here indisputable.

*Primary Parse 2: 49–**50–59**–61*. There is a major insertion at positions 54–59. The parse is confirmed by parsing strings within several insertions. For example, protein 36 has a GGGGGGG string at positions 53–59. The region of disrupted secondary structure extends strongly on the amino end of the deletion. Scattered P's are found starting at position 49. Residues are deleted as early as position 49 (in protein 44), and a PP sequence is found at position 53–54 in proteins 70–72 (MPI = 80%), and a PG at 50–51 (proteins 56–60, 63, MPI = 70%). The parse cannot be extended on the carboxyl terminus. The single deletion in protein 29 at position 60 is probably a misalignment; an R can be moved from position 51. There are scattered P's in positions 60–65, but not sufficient to extend the parse in this direction. Alignment anchors at positions 47 and 67 (APC E) make the alignment here secure.

*Primary Parse 3: 74–**75–80**–83*. There is a major insertion at positions 75–80. The parse is confirmed by parsing strings within the insertion. For example, protein 17 has the sequence GGRGPGG at positions 74–80. The parse might be weakly extended on the amino terminus by a deletion at position 74 in proteins 52–54 which is not easily realigned, but no further. There are scattered P at positions 81 and 82, and a widely distributed P at position 83. The deletions in protein 14 at positions 81–83 are probably an alignment error; a KPP sequence can be moved down from positions 76–78 moving the deletion to these positions.

*Primary Parse 4: 94–**95–100**–103*. There is a major insertion at positions 96–100. The parse is confirmed by parsing strings within the insertion. For example, protein 52 has the sequence PAGSN at positions 96–100. The parse might be weakly extended on the amino terminus by a deletion in only one protein at position 94 (a deletion that cannot be accounted for by an alignment error), but no further. There are scattered P at positions 101–103; this segment is PP in two proteins (67 and 68, MPI = 80%).

*Primary Parse 5: 111–**113–114**–114*. There is a 1–2 amino acid insertion at positions 113–114. The parse might be weakly extended on the amino terminus back to position 108, where the first P's are found. However, the P's do not appear to be significant before position 111. The parse cannot be extended on the carboxyl terminus. If the proteins containing the insertion are removed from the alignment, a secondary parse (a conserved GG string) would remain, thus confirming the parse.

*Primary Parse 6: 123–**126–128**–132*. There is a 3 amino acid insertion at positions 126–128. The parse is weakly confirmed by parsing strings within the insertion. For example, protein 36 has the sequence PPPG at positions 124–127. The parse might be extended on the amino terminus by a deletion at position 123, supported by P and G. However, as the preceding segment is assigned as a helix (*vide infra*), and a 3 amino acid deletion is found in proteins 39 and 40, this could indicate the loss of a single turn of a helix. There are scattered P at positions 129–132; the distribution does not make a strong case for an extension of the parse, however.

*Primary Parse 7: 167–**170—176**–176*. There is a major insertion at positions 170–176. The parse is only weakly confirmed, however, by structure disrupters within the insertion (a lone P in protein 81 at position 173). The parse might be weakly extended on the amino terminus by scattered P at positions 167–169. The sequences NG (proteins 5–8), PG (protein 26) and a deletion (protein 76) at position 169 make this extension at least plausible for a single position. However, the parsing element cannot be extended on the carboxyl side.

*Primary Parse 8: 190–**194–197**–200*. There is a 2–4 amino acid insertion at positions 194–197. The parse is weakly confirmed by P's scattered within the insertion, including a weak distributed parse in functional group 7 (MPI = 35%) (positions 196–197). The parse might be extended on the amino terminus by scattered P's back to position 190. Further, P's are scattered throughout the next segment leading to Primary Parse 9. Especially notable is a strong secondary parse at position 209, a conserved P. A PP sequence is found in proteins 46, 65, 80, and 84 (MPI < 35%). Thus, it is difficult to terminate this parsing region at the carboxyl end.

*Primary Parse 9: 213–**213–216**–220*. There is a major insertion at positions 213–216. The parse is not confirmed, however, by parsing strings within the insertion. The parse cannot be extended on the amino terminus. Thus, it is appropriate to examine the alignment in this region. The APC E (position 210) and CM3 P (position 209) firmly anchor the alignment on the amino end. The APC E is normally an indicator of an active site residue, as is the presence of conserved strings 4 positions (and longer) in length in subgroups with MPI values below 50%. The alignment is solidly anchored on the carboxyl side of this parse by positions 227, 228, and 229. There is a highly conserved P at position 209 which may indicate a secondary parse (see above). Scattered P's are found at position 218 and 220; none are found at 217 or 219. Thus, it is difficult to extend the parse in the carboxyl direction.

*Primary Parse 10: 223–**224–224**–225*. There is a 1 amino acid insertion at position 224 in proteins 28–31 (MPI = 50%). However, if the sequences of proteins 28–31 are shifted up by one position in the alignment between positions 217–224, the gap disappears. In its place remains a distributed parse extending from position 218–225 in the serine protein kinases (functional groups 1–7), but not in the tyrosine protein kinases. This parsing element is strong. A single deletion is found in protein 36 at position 220.

*Primary Parse 11: 241–**241–241**–241*. There is an insertion of a single amino acid at position 241. The parse is not confirmed by parsing strings in this insertion, nor is it confirmed on either side. The insertion might be an alignment problem; shifting the sequences of proteins 1–54 up by one position would move this deletion down to the major deletion that follows. However, the two plausible alignment anchors at positions 245 and 246 would be disrupted by this shift.

*Primary Parse 12: 244–**249–259**–263*. There is a major insertion at positions 249–259. The parse is confirmed by structure disrupters within the insertion. For example, protein 44 has the sequence PRGP at positions 250–253. The parse can be extended on the amino terminus by a deletion at positions 247–248, and scattered P back to position 244, with a highly conserved P at position 245. Widely scattered P's are present at positions 260–263, although the fact that the following segment is possibly an alpha helix could indicate that this segment involves the presence or absence of a single turn of a helix.

*Primary Parse 13: 272–**280–285**–289*. There is a single amino acid insertion at position 273. The parse is confirmed by 2 P's in the insertion. There are also 5 P's at position 272, and protein 36 has a PP sequence in this segment. On the carboxyl end, P's are scattered all of the way until position 276. Thus, it is difficult to end the parse. Position 275 may be a hydrophobic anchor for an external loop. Further, and most seriously, the parse might be collapsed by shifting the residues in functional subfamilies 1–4 up by one position.

For this discussion, the alignment has therefore been realigned to remove the deletion at position 273.

There is a major insertion at positions 280–285. The parse is confirmed by scattered P's within the insertion. For example, protein 26 has the sequence PWPS at positions 281–284. The parse is almost certainly extended on the amino terminus by deletions up to position 276; these cannot be accounted for by a misalignment. Further, P's are densely scattered back to position 276. Finally, disrupting sequences can be found in this region; a PPP is found in protein 42 at positions 276–278. On the carboxyl end, scattered P's are found from positions 286–289. However, the fact that the following segment is possibly an α helix could indicate that this segment involves the presence or absence of a single turn of a helix.

*Primary Parse 14: 314–**314–318**–320*. There is a major insertion at positions 314–318. The parse is only weakly confirmed by parsing strings within the insertion. The parse cannot be extended on the amino terminus. However, a few P's at position 320 might weakly extend the parse to this position.

It is worth noting that the end of this alignment does not correspond to the end of many of these proteins. It is, however, the end of the alignable catalytic domains. Thus, the end of the protein is not necessarily a strong parsing element in these cases. In functional subgroup 2, the sequence extends 12 positions before an apparent parse. An apparent parse is found in functional group 8 only two positions past the end of the alignment.

The regions of the alignment included in the deletion parses are assigned as surface coils, loops, and turns. Such structures are strongly indicated for the alignment numbers indicated in bold face. The extent of these structures is determined in part by the secondary structures assigned to the parsed segments that lie between them. These are considered below.

Secondary Structure by Segment

The primary parses listed above divide the alignment into 15 segments whose secondary structures are considered individually below.

The segments are designated by the alignment position numbers that they encompass. As noted above, the parses comprise different segments in different branches of the evolutionary tree. Normally, the ends of secondary structural units are expected to be different by one or two residues in different proteins. Therefore, the segments are designated by position number, with the bold face numbers inside the minimum extent of the segment, and the outside number (plain type) the greatest extension of the segment. This is the origin of the maximum and minimum lengths reported in Figure 5. The preferred length is tailored to the cAMP-dependent subfamily, as this is the subfamily presently being studied crystallographically. Normally, parses are included as the last position of a helix unless the parse is a deletion.

This variation in lengths presumably represents the fact that the structures of homologous proteins whose sequences have diverged by over 70%, although similar overall, are different in detail. This is, of course, a limitation on any method that builds models from a set of homologous proteins.

In this approach, a dialectic method is used. First, a 'canonical' assignment of secondary structure is made using simple rules. Other assignments are then said to have the burden of proof, meaning that the canonical designation is accepted unless the alternative designation is supported by the preponderance of evidence. In attempting to have other assignments meet this burden, the biological chemist is expected to construct as strong an argument as possible to set in opposition to the canonical assignment. This includes making *ad hoc* assumptions, restructuring the alignment, and questioning the assignments of positions to the surface and the interior of the protein. It is not important that this argument be reasonable. However, it should clearly indicate which canonical assignments must be set aside for the alternative assignment to be accepted.

Standard procedure in attempting to detect a helical segment involves mapping the surface/interior positions on a helical wheel. Should a pattern of amphiphilicity not be seen, positions are dropped from the ends of the segment, working towards the center, to determine whether the segment contains a subsegment that does display a 3.6 residue periodicity in conservation/variability. A canonical helix is assigned in this segment if amphiphilicity is seen over a segment 6 positions or longer. The effect of this procedure is to drop surface assignments at the ends of a helix when they fall in the 'inside' face of the helix; this is not irrational, as many positions on the hydrophobic side of an amphiphilic helix lie on the surface when they are at the end of the helical structure. Further, this procedure has the effect of dropping loops with hydrophobic anchors at the end of helical segments.

If the pattern remains non-amphiphilic, the helical wheel is examined to identify assignments that, if changed, would create an amphiphilic helix. This provides a set of *ad hoc* assumptions (e.g., the helical assignment can be made if the assignment of position x to the surface is incorrect). Then the procedure is begun again, but this time beginning with the strongest assignments and working towards the weakest. The goal of this process is to find the strongest possible argument that the segment is helical. Quite often, the segment will be assigned to some other structure canonically, and this argument will serve only to establish a dialectic with respect to the canonical assignment.

Questioning the assignments involves several types of arguments.

Figure 3 lists every position where more than one subgroup is variable at a MPI > 50% where more than 1 of the variable subgroups has at least one polar residue (D, E, K, N or R). For an alignment as large as protein kinases, this produces assignments which range from extremely strong to extremely weak. For smaller alignments (e.g., ADH, with only 17 proteins), the analogous table was constructed so that surface assignments are made where more than one subgroup is variable at a MPI > 50% where at least 1 of the variable subgroups has at least one polar residue. The stronger assignments are those with more variable subgroups in clusters with higher MPI values, with more of the variable subgroups containing polar residues.

There are many reasons why assignments might be incorrect. For example, in multimeric proteins, the most common interior misassignments are positions that lie on the surface of a subunit but form a subunit–subunit contact. The biological chemist must be aware of these complications when building models. Indeed, in ADH, such ambiguities can be used to advantage, as subgroups within the protein family are dimers while others are tetramers, and contact sites can be predicted based on patterns of sequence divergence and conservation (see Reference 4). Analogous problems are also expected when the protein interacts with a membrane, a real possibility with the protein kinases.

Paradoxical assignments are straightforward to identify from Figure 3. Typical is a position with a large number of variable subgroups with only a few containing a polar residue. For example, position 135 has 10 variable subgroups in the cluster of subgroups with MPI = 60%, yet only 2 of these have polar amino acids. This can be contrasted with most other surface assignments, for example the nearby position 125, also with 10 variable subgroups at this MPI, but with 8 of the variable subgroups containing polar residues.

Such discrepancies trigger closer examination of the alignment at this position. Misassignments can arise, of course, from misalignments, where a hydrophobic variable position in several proteins (a strong indicator of an interior position, see Reference 5) is misaligned to match with a surface position in other proteins. Alternatively, the discrepancy is expected when members of the protein family have different folded structures. Closer inspection of position 135, for example, shows that function groups 7–12 all have a hydrophobic residue. The surface assignment is made entirely based on structural variation in functional groups 1 and 3–6.

A weak pattern of amphiphilicity can be further tested by breaking the alignment into subgroups and looking for segments that display more strongly a 3.6 residue periodicity. Helices whose amphiphilicity is obscured by misalignment generally become obvious by this procedure (see for example, segment 115–**115**–**122**–125).

Internal helices are less frequent than surface helices and, because they do not have strong surface assignments, are more difficult to identify in the first phase of the analysis. The markers for an internal helix are a segment devoid of parsing elements too long to form a single beta strand given the assumed dimensions of the folded protein, patterns of amphiphilicity at the ends of the structure, and 3.6 residue periodicity of variability of (normally hydrophobic) amino acids within the segment. For example, the two largely internal helices of ADH are found in this way. The longer passes near the active site, where the 3.6 residue periodicity is quite obvious.

When a putative helix is found, it can be confirmed in several ways. Most convincing is to find a subgroup of proteins in the alignment where the orientation of the helix relative to the globular protein appears to be shifted, and the inside/outside assignments for the subgroup appear to have moved 20–30° around the helical cylinder. The helix in segment 115–**115**–**122**–125 is assigned strongly this way. Second, co-variation analysis of segments of a helix often detects pairs of positively and negatively charged amino acids 3 or 4 positions apart that appear and disappear together during divergent evolution, suggesting an intrahelical contact.

The standard procedure for assigning a beta strand is to map a segment on an alternating template. Beta strands are rarely as obvious as alpha helices by this process when the entire alignment (with MPI < 40%) is examined. Therefore, several expedients are adopted.

First, beta strands are often internal structures, and therefore appear as strings 2–8 positions long with consecutive interior assignments. A hydrophobic stretch of this length is canonically assigned as an interior beta strand. Further, the pattern of variability of a putative internal beta segment is apparent as a progression from less conserved to more conserved to less conserved, with this pattern superimposed on an alternating pattern. Such structures are normally highly ordered, with splits at MPI > 50%, and are not easily mistaken for surface helices or coils and loops.

Nevertheless, two major ambiguities remain in our approach in assigning beta strands. First, very short beta segments (2 amino acids) are generally difficult to assign correctly. Second, surface beta strands are often confused with surface coils. The latter may be an intrinsic limitation of any method that builds structural models from a set of homologous proteins, as the conformation of surface beta strands is not necessarily conserved during divergent evolution.

The most important procedures for identifying loops and coils are discussed under parsing elements. Alternatively, a string of four consecutive surface assignments is canonically assigned as a surface loop, an assignment that is rationalized by the classical argument that four consecutive polar amino acids in a single protein sequence indicate a surface loop or turn (17). Often a surface loop is observed to have a single position where a hydrophobic amino acid is found (often hydrophobic variable). This is termed a 'hydrophobic anchor for the loop, and they serve as indicators of short loops.

Units 1 and 2. The first segment can be divided into two subsegments based on the exposure to the surface, the first from positions 1–7 (6 out of the 7 positions are assigned to the surface), and the second from 8–24, containing only 3 highly variable surface positions (with the surface assignment at position 11 weak due to the small number of variable subgroups having polar residues), three intermediate assignments (positions 8, 16 and 19), and 5 internal assignments, out of a total of 17 positions. The alignment is well anchored in these positions, with APC residues (or nearly APC residues) at positions 10, 12, 15, and 17.

When dividing long segments, it is customary to look for a secondary parse that might break the long segments into two or more shorter segments. The region has scattered prolines (at positions 5, 7, 8, 11, 23 and 24). None of these create a strong secondary parsing element as, consistent with the philosophy of our method, they may be prolines introduced into a standard secondary structural unit to engineer in a desired amount of instability into a standard secondary structural element.

The APC G at position 12 is a stronger secondary parse. Glycines that are absolutely conserved in an alignment with a minimum pairwise identity of <35% often divide secondary structural units (for example, in ADH, amino acids 66, 71, 86, 192, 201, 260, 261 and 320, but see in contrast position 204, an APC G that is in the first turn of an alpha helix). This suggests a division at positions 10 or 12. There are some strings that indicate a stronger parse in this region. For example, proteins 67–69 conserve (in a subgroup with an MPI = 70%) a GGG string at positions 10–12. GG strings are found in proteins 9, 46, and 52–55. More interesting is the fact that in protein 44, the T at position 10 that substitutes for an otherwise completely conserved G is complemented by a P at position 11.

Regions, such as subsegment 1, with a high fraction of surface assignments canonically are designated as coils or loops; the canonical assignment for segment 1–7 is therefore a coil. It should be noted that in most of the proteins examined here, position 1 is not the first in a separate polypeptide chain, but is normally fused to a transmembrane region, from which it can be released by proteolysis. Such proteolysis often occurs in coils or loops, providing independent support for the canonical assignment. In the cAMP-dependent kinases specifically, the polypeptide chain begins 30–40 amino acids before the beginning of the alignment. It is conceivable that this leading segment adopts a standard secondary structure. If so, this structure is almost certainly broken by a distributed parse at positions –7 to –5 (where the alignment in Figure 1 starts at position 1). However, there is little evidence from the sequence of this leading segment that a standard secondary structure is adopted.

Dialectically, an amphiphilic helix can be built from position 1–14 only if position 8 is regarded as a surface (K = 5, 6 variable subgroups, 5 with polar residues), the surface assignment at position 2 is ignored (justified by the generalization that positions 'inside' but at the ends of an amphiphilic helix often lie on the surface), and the surface assignments at positions 5 and 6 are treated as weak (only 2 of 7, and 4 of 13 variable subgroups have polar residues). These *ad hoc* assumptions are weak. For example, the surface assignments for position 6 (13 variable subgroups, 4 with polar residues) and position 11 (10 variable subgroups, 5 with polar residues) are of essentially identical strength. Yet in the amphiphilic helix, position 6 is in the middle of the 'interior' region, while position 8 is in the middle of the 'surface' face of the helical projection. On these grounds, the assignment as a coil is preferred.

Concerning the second subsegment, regions with a low fraction of surface assignments, a large number of splits, and no pattern of helical amphiphilicity are canonically designated as beta strands. Thus, the canonical assignment of the second subsegment is as an interior beta strand. The pair of surface assignments at positions 21 and 23 suggests that the beta strand continues until the very end of this subsegment. The second subsegment is largely inside, and hydrophobic. Further, the strongest surface assignments are at positions 19, 21, and 23, an alternating pattern that indicates canonically that this is a beta structure. Essentially no evidence suggests that this segment is an internal helix.

Even in its shortened version (positions 12–22, 11 amino acids), this beta strand is longer than typically found in folded proteins (18), and rather long considering the small size of the protein overall. The strand can be shortened (Figure 5 suggests that the 8-position strand from position 14 to 21 is the minimum length). Alternatively, the beta strand can be conveniently bent at position 15–16 (secondary parse) to form two shorter beta strands, the first 4 positions in length, the second 7. At a later stage, when the secondary structural units must be assembled into a globular structure, such bends are important.

We have no grounds for preferring a start of the strand at position 12 or 13. The choice of the end point at position 22 is based on the alternating pattern observed in some subgroups of the alignment, and the need to have a sufficient number of filled positions in functional subfamily 7 to execute a standard turn before the next secondary structural unit (*vide infra*).

This secondary assignment, classed as 'strong', is likely to be controversial. This segment contains the string GXGXXG, well known in proteins that bind ATP and adenine dinucleotides (19). Thus, several investigators have suggested that this segment indicates that the catalytic domain of protein kinases is homologous to other kinases and, in particular, this first region is a strand-turn-helix structure. A particularly convincing case for this is made by Sternberg and Taylor, who compared this region with the NAD+ binding regions of glyceraldehyde-3-phosphate, lactate and alcohol dehydrogenase, and the FAD binding regions of glutathione reductase and p-hydroxybenzoate hydroxylase (20). In these proteins, the GXGXXG sequence lies between a beta strand and an alpha helix (a beta-turn-alpha structure), and Sternberg and Taylor concluded that this region had a similar conformation in the protein kinases discussed here.

Unfortunately, it has not been possible to find evidence using our method for a beta-turn-alpha structure for the segment 1–26. Indeed, the canonical assignments are for a coil-turn-beta structure, and we have little grounds to revise these assignments.

Therefore, we conclude that the GXGXXG motif found in the protein kinase alignment does not mean that the secondary structure preceding and following this motif is the same as in other proteins where this motif is found. Further, we do not believe that this particular motif is sufficient evidence to conclude that the dehydrogenases, reductases, and hydroxylases mentioned above are homologs of the catalytic subunit of these kinases.

There are no assignments of active site residues in this segment. The absence of active site assignments in this region contradicts an experimental fact, that Lys 7 is protected from modification by acetic anhydride in the presence of Mg–ATP and an inhibitory peptide; which may indicate that this segment is near the active site. It will be interesting to see the position of this residue as determined by the crystallographic work (21).

Unit 3. The proteins of functional subfamily 7 have a gap in the alignment that extends fully from position 25 to position 41. This fact implies that positions 24 and 42 lie close together in space in the other proteins, suggesting that if the beta strands assigned to positions 12–22 and positions 42–48 are part of the same beta sheet, they lie antiparallel. Further, it leads us to terminate the preceding beta strand one residue sooner and start the following beta strand one position later than would be done based on analysis of the first segment alone. There is no evidence for active site residues in the coil.

Unit 4. The dominant features of this segment are the interior assignments to positions 43 to 46 and 48. Further, there are several hydrophobic splits throughout the alignment, indicating an ordered structure (rather than a coil). The canonical assignment for this segment is a beta strand. This assignment is strongly confirmed by the two-position periodicity in the pattern of conservation and variation, especially evident in subgroups of the alignment with MPI values near 60%. For example, functional group 11 (MPI = 50%) has the string 43 (V or C); 44 (conserved A); 45 (I or V); 46 (conserved K); 47 (surface); 48 (V or L); 49 (surface). Thus, this secondary assignment is classed as 'very strong'.

The most significant evidence for a 3.6 residue periodicity that might suggest an internal helix is the 3 amino acid deletion at the beginning of the segment (positions 39–41) in functional subfamily 7, and the deletion of three amino acids at positions 49–52 with respect to protein 29, and four amino acids (49–53) in protein 44. The strong alignment anchors at positions 44 and 46 fix this deletion. However, these deletions are also consistent with a coil structure at the ends of this segment, and no other evidence appears to suggest a helical assignment for this region.

Position 46 (APC K) is strongly assigned to the active site. This assignment is consistent with experimental data. Lys 46 is modified by fluorosulfonylbenzoyladenine, a substrate analog. Further, when cAMP-dependent protein kinase is treated with dicyclohexylcarbodiimide, it appears that a major product is an intrachain crosslink with Asp 182. Both residues are absolutely conserved, and the modification is blocked by the presence of Mg–ATP (22).

Unit 5. This segment contains both deletions and parsing elements, and therefore is assigned a coil structure. Protein 44 has the entire region from position 49–59 deleted, implying that positions 48 and 60 lie close together in space in the other proteins. Functional subfamily 8 also deletes much of this region, with parsing indicators present as late as position 63. Functional subfamily 10 (MPI = 65%) shows co-variation between positions 50 (EEK) and 55 (KKE), consistent with an interaction between these two residues in this subgroup of proteins. The parse in this region has some noteworthy features. Protein 36 has a string of 7 consecutive glycines (positions 53–59). Also, there are 3 consecutive D's in functional subfamily 2 (positions 56–60). The scattered P's at positions 60–63 suggest that the parse might extend as far as position 63. There is no evidence for active site residues in the coil. However, the segment 52–57 could adopt a standard secondary structure in functional subfamilies 1 and 2.

Unit 6. The alignment is well-anchored in this region by the APC E at position 67. On the amino end, the alignment is secure at position 46 (APC K).

From position 64 to 73, the interior and surface assignments map well on a helical wheel (Fig. 4a). Therefore, this segment is canonically assigned as a helix. This assignment is initially classed 'moderately strong', as it involves relatively weak surface assignments. Indeed, the strongest surface assignment based on variable subgroups is at position 65 (9 variable subgroups, K = 4). However, only 3 of the variable subgroups at this position have a polar residue (DEKRN). Indeed, the amino acids found at position 65 are largely hydrophobic, and it is also worthy of note that proteins 46 and 61 have P at this position. Position 66 has the largest number of variable subgroups (8) where the variable subgroup also includes a polar residue (7). Position 64 is a functional variable (7 variable subgroups), a strong indicator of an inside position in an alignment this large. This position is also listed as a hydrophobic variable (4 variable subgroups, K = 5). Position 65 is listed as a hydrophobic variable with 2 variable subgroups (K = 9,8). The rather weak assignments of positions to the surface suggest that this helix is partly buried, especially when compared with other helices (e.g., the helical conformation assigned to unit 14).

To strengthen (or possibly weaken) the assignment, other indicators are examined. Co-variation analysis identifies a few interesting residues that are consistent with a helical assignment in this position. For example, in the three proteins in functional subfamily 10, the residues at position 62 are KKE, while the residues at position 66 are DDR. This pattern of co-variation is consistent with an intrahelical contact between the side chains at these positions in these proteins, and provides further support for extending this helix to position 61 in at least some of the functional subfamilies.

The first four positions of this segment (positions 60–63) are assigned to the surface, canonically indicating a coil. Scattered P's in this segment (e.g., conserved in proteins 33–36, MPI = 50%, 56–59, MPI = 80%) and the parsing string PXP (protein 62) confirm this assignment as a coil. However, a dialectic position is possible that the helix extends another full turn, starting from position 60. In this extra turn, position 63 falls on the inside of the helical projection. As this position lies at the end of the helix, this assignment is not entirely incompatible with the moderate surface assignment at this position (Fig. 3). However, it is conceivable that the first turn of the helix is missing in proteins in functional subfamily 8 (for example), where the preceding coil is short.

In the preferred assignment (Fig. 5), the last position is position 74 due to the inside assignment. The choice of position 61 as the first position of the helix is somewhat arbitrary; it was made because the cAMP-dependent protein kinases have a hydrophobic residue at this position, and the parses at positions 60 and 75 match.

Especially noteworthy about this segment is the absolutely conserved E at position 67. This is a glutamate in the middle of the hydrophobic region. Canonically, one assigns an APC E as an active site residue, which would place this helix at the active site. There are, of course, other roles for amino acids with functionalized side chains that are sufficiently important that the residue at that position is highly conserved. For example, in lactate dehydrogenase, an APC E (311) is found in the middle of the final helix where it forms a hydrogen bond to W 203 (which is also conserved). Likewise, an APC E in alcohol dehydrogenase is found at position 35 in a beta strand not at the active site. In the putative helix in the protein kinase family, there is little to indicate that this segment is at the active site; for example, there is no string of conserved residues at low MPI (compare, for example, the pattern of conservation in the active site helix between positions 46 and 55 in the ADH family). Thus, for future assembly of the secondary structural elements into a folded structure, this helix need not end up at the active site.

In this context, several experimental facts are valuable. First, Glu 67 is especially reactive with carbodiimides, implying that it lies in a hydrophobic region, and the reactivity is diminished in the presence of Mg–ATP (23), implying that it lies near the active site. Further, the lysine at position 68 appears to be inside the protein fold based on its slow reaction with acetic anhydride (20).

Unit 7. In functional subfamily 7, all amino acids are deleted from position 74 to 80, indicating that positions 73 and 81 are close in space in the remaining proteins. There is no evidence for active site residues in the coil.

Unit 8. This is the first of several beta-like segments that are broken by secondary parsing units at several places throughout. Only two strong surface assignments are made after position 84 in this segment, at positions 87 and 94. A third, at position 90, is weak enough to be indecisive. The remaining positions are hydrophobic splits (e.g., at positions 86 and 90, the splits are perfect at MPI = 70%). Thus, the canonical assignment for this segment is a beta strand from position 84 to position 93. However, the assignment is classified only as 'strong' because it is difficult to confirm this assignment by alternating patterns in the alignment.

Central to this difficulty is the fact that prolines are found scattered throughout this segment, at position 87, and at position 91. These secondary parses are potentially valuable as the beta strand is long, and is probably bent. Position 87 is interesting in this regard. Positions displaying variability of this sort are often found in beta strands that emerge above the surface of the globular protein at a point where the beta strand bends and then continues, in this case until position 93 or 94. The fact that isolated P's are found in proteins 40, 44, 50, and 57 at position 87 supports this picture, as does co-variation analysis with the following segment (*vide infra*). It is worth noting that the lysines at position 87 and 93 appear to be inside the protein fold based on reaction with acetic anhydride (24). Finally, the carboxyl side chain of E 89 is especially reactive with carbodiimide, and is not protected by substrate, suggesting that it is inside the folded structure but not at the active site.

The parse at the amino end extends strongly to position 83, suggesting that any beta-like structure begins only at position 84. The segment is ended at position 93 in the cAMP-dependent protein kinases (Fig. 5) because of the deletion in protein 39. The two-amino acid deletion could, however, indicate that in the remaining proteins, the beta strand continues two positions further.

Dialectically, one can attempt to assign this segment a helical structure. Positions 84, 87, and 94 can be assigned to the surface with varying degrees of reliability. These map on one side of a helical wheel. Position 90 maps extremely weakly on the surface, and appears on the correct side of the helix. Position 91 violates the amphiphilic helical pattern. Based on the assignments at positions 90 and 91, it is unlikely that this segment is a surface helix; however, a helical assignment might be made if the assignments at positions 90 and 91 were in error. Therefore, the possibility that it is an internal helix, or a helix forming a contact with a membrane or another protein subunit cannot be ruled out.

However, 'functional variable' assignments at positions 85, 86, 91, and 92 are distributed evenly around the helical projection, and offer no support for a helical assignment. Further, none of the 12 helical wheels for the functional subfamilies show convincing patterns of amphiphilicity, either in the pattern of variation or in the polarity of the amino acid residues themselves, largely because position 86 remains firmly inside. Thus, a helical assignment is rejected.

Unit 9. Positions 94–100 contain no amino acids in protein 39, suggesting that positions 93 and 99 are close in space in the remaining proteins. However, protein 39 is among the most structurally divergent from the bulk of the alignment, and a more reliable hypothesis is that positions 95 and 99 are close together in the folded structures. This implies that if the preceding and following beta strands belong to the same beta sheet, they reside antiparallel. Co-variation analysis (*vide infra*) suggests that this is the case. There is no evidence for active site residues in the coil.

Unit 10. This segment appears to be another long broken beta strand. The alignment is anchored in this region at position 108. Especially unusual is the 2 residue deletion (positions 108 and 109) in protein 5, which is 90% sequence identical with proteins 6 and 8. This deletion corresponds to a single P in functional group 11 (protein 85), and two P's in functional group 6 in proteins 40 and 44. Position 108 is unusual; it is a highly conserved E in the middle of a hydrophobic stretch. This deletion can be moved down in the alignment by realigning the segment; the alignment anchors in this region are not decisive, and this realignment is the basis for the analysis that follows.

The segment forms part of the interior core of the protein, with positions

103–110 all assigned to the inside (position 105 has 5 variable subgroups, K = 4, with a single variable subgroup with a polar residue). This is canonically assigned as a beta strand.

Again, the beta strand is rather long, and probably bent at position 108. This notion has an interesting consequence in the context of the assignment of unit 8. There is substantial co-variation between position 87, which has a high proportion of basic residues, and position 108, which has a high proportion of acidic residues. In particular, in proteins 40, 44 and 85, the E at position 108 is replaced by a P, and in all three cases, the basic residue at position 87 is replaced either by a P (proteins 40 and 44) or by an S (protein 85). The co-variation is striking in functional group 7, where position 87 has LLPLRRR matched against QQQQEEE at position 108. Thus, it seems to be a reasonable working hypothesis that these beta strands are in close proximity, with the bend at position 87 matched to the break at position 108, with a contact formed between the side chains of the amino acids at these two positions.

Co-variation analysis is one of the most problematic ways of finding contacts between distant points in a polypeptide chain. We do not yet have programs to do co-variation analysis on an alignment as large as protein kinase. Further, co-variation is rarely perfect. However, in the absence of disulfide bonds, it is the only way to constrain the distance between portions of the chain that do not lie at the active site. In any case, it seems certain that co-variation analysis is most useful when one already has a structural hypothesis, as is the case here.

Unit 11. This coil has little variability in length, with 2–4 positions. There is some experimental evidence that this coil lies near the active site. When the phosphorylatable serine in the peptide LRRASLG is replaced by the photoaffinity agent, *p*-benzoylphenylalanine, Gly 113 and Met 116 are modified (25).

Unit 12. Positions 115–123 of this segment show a 3.6 residue periodicity indicative of a helix (Fig. 4b). However, the assignment is classified only as 'strong'. A plot of the surface/interior assignments on a helical wheel shows only a weak amphiphilicity. Positions 115, 117 and 119 are assigned only weakly to the surface (only 5 variable subgroups K = 4). The strongest surface assignment is at position 118, and at 122, but here the amino acid is at the end of a secondary structural element (given the parse that follows), making it unlikely that a surface assignment here would be definitive in ruling in or out a particular assignment.

As shown below, there is reason to believe that the serine and tyrosine kinases are not perfectly aligned in this region. A misalignment has two implications. First, it is difficult to identify secondary structure from an examination of the entire alignment. Second, examining fragments of the alignment should be a more productive way to assign secondary structure.

The segment is flanked on both sides by deletion/insertions of varying lengths. Second, there are no anchors for the alignment within the segment. Indeed, there are essentially no anchors for some distance outside the segment. The Gly at position 113 might be considered as an alignment anchor, but it comes in a loop region. The substitutions at position 106 provide a weak anchor, as might the scattered P in the two main classes of kinases at position 132. However, on the amino end of the segment, the last completely solid anchor is at position 67, although weaker anchors might be found at position 108 (where the ancestral residue is presumably E) and perhaps the hydrophobic element at positions 103–107 (although alignments based on such an orientation can be plus or minus a single residue). The next completely solid anchor on the carboxyl side comes at 155–157 to the carboxyl side, although weaker anchors exist at positions 144, 140 (where Q might be reconstructed as the primitive residue in the common ancestral sequence), and 139 (with A as the common ancestor). Clearly, in between, the alignment is unreliable.

However, when the segment is divided into the 12 functional classes and the data replotted, the helical ambiguity goes away. For example, functional subfamily 3 at position 119 has either the hydrophilic D (proteins 20 and 21) or R (proteins 24, 25, and 26), while the position is the hydrophobic Y in proteins 22 and 23. The distribution is tree-like based on the overall sequence identities of the proteins. The situation is mirrored on the other side of the putative helix. Position 121 has either the hydrophobic V (proteins 20, 21 and 24) or I (proteins 25 and 26), while the position is the more hydrophilic T in proteins 22 and 23. Thus, it appears that in proteins 22 and 23 the helix is rotated a bit 'clockwise' with respect to the folded structure of the protein. Table 1 lists the borders of the helix in the different functional subfamilies.

This region illustrates also how information appears at different levels of sequence divergence. Group 2 does not obviously show an amphiphilic helix, and a helical assignment of this segment would not be secure if these were the only sequences available to the biological chemist.

When a helix has become 'reoriented' by a residue (a statement that, at the present level of analysis, is equivalent to a statement that the alignment is misconstructed due to a shift of one position), the assignments made across the shift become confused, and assignment of secondary structural elements becomes weaker. However, if a helical pattern is strengthened by assignments based on a partitioning of the alignment to avoid including shifted (or misaligned) structures within a single alignment, a 'very strong' assignment of secondary structure can be obtained.

For the cAMP-dependent kinases, the preferred helix extends from position 115 to position 124. The fact that position 123, assigned to the surface, appears on the inside face of the helix is acceptable as it is at the end of the helix. The last turn of the helix is probably missing in members of functional subfamily 6.

TABLE 1. ORIENTATION OF THE HELIX BETWEEN POSITIONS 115 AND 122

Functional subgroup	MPI	Outside-inside border	Inside-outside border	
1	50%	115–119	117–121	
2	80%	119–123	118–122	Largely buried? MPI = 80%
3	45%	119–116	121–118	Note shift in 2 proteins
4	50%	115–119	121–118	
5	45%	115–119	120–117	
6	30%	115–119	117	117 Hydrophobic and -philic
7	35%	115–119	121–118	
8	60%	115–119	117–121	
9	50%	115–119	117–121	
10	65%	115–119	117–121	
11	50%	115–119	117	117 Hydrophobic and -philic
12	55%	115–119	117	117 Hydrophobic and -philic

Unit 13. In proteins 39 and 40, positions 123–128 are deleted, suggesting that positions 127 and 129 are close in space in the remaining proteins. Further, this suggests that the parse extends until position 132, although it can be argued (*vide infra*) that the following helix begins as early as position 131. There is no evidence for active site residues in the coil.

Units 14–17. This is the longest unparsed segment in the alignment (41 positions with no insertions or deletions). A helix of this length would have 11 turns, and be approximately 60 Å in length; a single beta strand would, of course, be longer. 60 Å is considerably longer than the expected diameter of a spherical globular protein of this length, implying that there should be an internal parse. The first four positions (129–132) are plausibly coil structures (note the parsing string PXP conserved in part of functional group 1, MPI = 80%). A secondary parse is present at position 160. These potentially divide or shorten standard secondary structures that might be assigned to this segment.

However, the most striking features of this segment are the highly conserved strings RDLK and RDLA at positions 155–157. Such strings are assigned canonically to the active site, and are often found in loops or coils, making plausible a break in a helical structure up to this point. Further, the reaction of E 161 with water-soluble carbodiimide is partially by Mg–ATP, and fully blocked in the presence of both Mg–ATP and peptide substrate (26), experimental evidence favoring an active site assignment for this segment.

The segment has the most distinctive alpha helical pattern in the protein, with amphiphilicity extending in a 'textbook' fashion from position 131 until 151, over 5 turns (Fig. 4c). Textbook surface helices of this length are essentially never misassigned, making this among the strongest secondary assignments in this alignment.

It is worth a few words to examine the extent to which this helix is confirmed by the details of the pattern of conservation and variation. Position 135 at the boundary between the surface and inside faces of the helical projection is rather weakly assigned to the surface. Although there are 10 variable subgroups at MPI = 60%, only 2 contain a polar amino acid. Closer inspection of the alignment shows that the surface assignment comes from subgroups in the serine/threonine kinase functional class; the sequences of the tyrosine kinases alone would yield an interior assignment.

It then remains to determine how far this helix can be extended in either direction. The amphiphilic pattern is disrupted in the amino direction by the surface assignment at position 130 (which is on the inside face of the helix) and by the inside assignment at position 129 (which is on the surface face of the alignment). Position 130 may still be a part of the helix, of course; as noted previously, 'inside' amino acids at the end of amphiphilic helices are often found on the surface. However, this implies that the helix does not extend more than a half turn further in the amino direction. This notion is confirmed by the scattered P's at position 132. Again, P can be in a helix, but normally within one turn of the amino end. Based on this reasoning, the amino terminus of the helix is designated as position 131.

The carboxyl end of the helix is more difficult to assign. Position 149 technically breaks the amphiphilic pattern of the helix, but with only 3 variable subgroups (MPI = 50%), and only 2 of these having polar residues, this is a weak surface assignment and is not decisive. Positions 150 and 151 are assigned to the surface and appear correctly on the helical wheel. Position 152 is a surface assignment appearing on the 'inside face' of the projection, 153 an 'inside' assignment on the inside face, and position 154 an inside assignment on the surface face of the helix. These positions mark the end of the amphiphilicity in the helix.

Attempts to extend the helix further are problematic, as the segment now becomes highly conserved and highly functionalized. Consecutive conserved functional residues are strongly assigned to the active site, and functional constraints and adaptive variation often obscure patterns that would otherwise be good indicators of secondary structure.

The helix is rather long (22 amino acids, a bit over 6 turns), with a total length of ca.33 Å. Thus, it traverses the entire globular structure of the protein, and it is difficult to imagine the active site perched on the end should we extend the helix much farther than position 153. The implication is that the helix must end here, that the parse at position 160 must be a real break in secondary structure, and that the secondary structure of the segment that follows (from position 161–169) must be considered separately.

The remainder of the segment is divided into three units. The first (unit 15 in Figure 5) reflects the possibility that there is a short beta strand at the end of the helix (positions 153–156). This canonical assignment is based on the large number of splits. Unfortunately, this assignment is classified as 'weak' because splits near an active site are also indicative of coil structures.

The next subsegment (unit 16) is clearly an active site coil based on the APC D at position 157 and the APC N at position 162.

The following subsegment (unit 17), a string (positions 163–165) of interior assignments, is canonically assigned as a beta strand.

The lysine at position 159 appears to be inside the protein fold based on reaction with acetic anhydride. Further, the carboxyl side chain of E 161 is especially reactive with water-soluble carbodiimide; it is completely protected by Mg–ATP and an inhibitory peptide. This confirms the assignment of this region to the active site (23).

Unit 18. The four consecutive surface positions (166–169) at the end of the previous segment are canonically assigned as a coil. As the deletion in protein 76 starts at position 169, and the alignment is well anchored in this region, the minimum length of the parsing segment is 2 amino acids, and position 168 is likely to be near position 177 in the remaining proteins. There is no evidence for active site residues in the coil.

Units 19 and 20. The conformation of this segment is the most difficult to assign in the entire alignment. Problems arise because within this segment differences in the sequences of the different subgroups of the kinases make the alignment difficult to construct reliably. However, these differences cannot be explained as the random variation expected for a segment that lies far from the active site as there is an absolutely conserved triplet DFG at positions 182–184. The conserved triplet is an extremely strong indicator of a segment near the active site.

It is worth noting here that it is the conserved *string* of three residues that makes this assignment strong. APC D is known at positions other than the active site, and APC G is often simply a parse. APC F is occasionally found in the center of hydrophobic cores. Thus, each conserved amino acid alone would not be a strong indicator that this segment lies near the active site. Further, nearby positions display reflexivity. For example, at position 186, functional subfamilies 1–7 have A, S, and C, while functional subfamilies

8–12 have A, S, and T. Such reflexivity often indicates a purely structural constraint on divergence. However, a conserved string such as the one present here is essentially never found far from the active site.

Chemical modification studies also suggest that this segment is at the active site. Treatment of a cyclic AMP-dependent protein kinase with carbodiimide yields a protein cross-linked between the side chains of Asp 182 and Lys 46 (27). Mg–ATP blocks this reaction. Chemical modification studies support the notion that K 46 lies at the active site, implying that the pair is at the active site.

Extreme variation in sequence near an active site is, of course, anticipated for a set of homologous proteins whose function has undergone divergent evolution. This is a strength of the approach used here, as neither variation nor hydrophilicity is automatically assigned to the surface, but rather only specific types of variation and hydrophilicity. Thus, the variation in this region is interpreted as evidence that this segment forms a structure at the active site that is important for binding the protein substrate. As the structure of the natural protein substrate is quite different for different functional subfamilies, one expects both the sequence and the folded structure of this region to be different in different proteins, meaning that the assumption central to our approach is probably not entirely true in this region of the alignment.

Canonically, a string of inside residues and splits (e.g., positions 177–186; note that the surface assignment at position 177 is weak) is assigned a beta strand structure. Here, it seems plausible to associate the strand at position 182 with the beta strand assigned between positions 42 and 48, primarily on experimental data obtained by cross-linking experiments. As discussed below, these two beta strands appear to be aligned antiparallel to each other. Further, if the first beta strand is bent at position 46, so would the second (at position 182).

Positions 190–193 are assigned to the surface, position 90 strongly in all subgroups of kinases. Four consecutive surface assignments are canonically assigned as a coil. In functional subfamilies 1–6, these positions also contain scattered P's throughout. Thus, the canonical assignment of this subsegment is a coil in this subfamily. However, in other functional subfamilies, other secondary structural assignments are possible, especially as no P's are found at this position in functional subfamilies 7–12. It is important to look closely at these, starting with the functional subfamily with the *lowest* MPI, as this is the subgroup that retains the most information, provided that the tree within is appropriately branched.

Functional subfamily 7 has a MPI = 35%, and is divided into two subgroups (proteins 48–51 and 52–54, respectively) with MPI = 80%. Thus, the extent and distribution of diversity of the proteins in functional subfamily 7 is nearly satisfactory for this functional subfamily to serve (by itself) as the basis for an application of our structural prediction method. Indeed, did we not have the other sequences, we would attempt to assign a structure from these 7 proteins alone, although the reliability of the structural assignments would, of course, be much lower than the ones presented here.

The segment in functional subfamily 7 between positions 184 (APC G, viewed as a parse for this discussion) and 191 maps as an amphiphilic helix. At positions 188 and 189, the apparently contradictory assignments (surface and inside) in fact divide exactly according to the two subgroups of functional subfamily 7. Proteins 48–51 are 'inside' at position 188 and 'outside' at position 189, while proteins 52–54 are 'outside' at position 188 and 'inside' at position 189. This suggests that the helix is turned slightly in proteins 48–51 in comparison with proteins 52–54, and the pattern strengthens what would otherwise be a weak helical assignment. The segment also contains charge variation at position 191 (R or D) and at position 188 (V or K); this could be co-variation indicating a helical structure, but analogous indicators in proteins with known structures are not highly reliable. Nevertheless, the assignment of a helix in this region is satisfactory for us to make it the preferred assignment in this region for this subfamily of proteins.

Two functional subfamilies, 3 and 5, have MPI = 45%. In functional subfamily 3, the segment 185–190 could map as an amphiphilic helix, but the clarity of the map is compromised by uncertain assignments at positions 186 and 187. In functional subfamily 5, a helix is possible from positions 185–189, but the number of positions assigned is too small for the amphiphilicity to be significant.

Functional subfamily 1 has a higher MPI (50%), meaning that still more information is lost in the subalignment. Nevertheless, a pattern of amphiphilicity resembling a helix from position 185 to position 189, and possibly to position 193, is observed. The parse at position 190 is preferred based on a comparison with other subgroups, but is not absolute.

It is worth noting that surface assignments made for these subfamilies are weaker than those made when the entire alignment is considered. For example, position 187 has an APC K (MPI = 50%). With a subalignment of so few proteins and such little overall divergence in sequence, it is difficult to tell whether the conservation indicates a functional constraint on drift, or whether it merely indicates that insufficient sequence divergence has taken place for us to have (fortuitously) found a protein where the codon corresponding to this position had undergone mutation. For our purposes here, the position has been assigned as the weakest surface, simply based on the polarity. However, it is worth noting that K 187 is less reactive with acetic anhydride than other lysines in the protein, and therefore appears to be inside the protein fold.

To search for secondary patterns of conservation that might indicate that the active site segment exists in a particular standard secondary structure, the number of conserved subgroups in clusters of subgroups with different minimum pairwise were examined (Table 2). The data show a general trend; conservation falls off in both directions as one proceeds away from positions 182–184. However, in the amino direction, there is a pronounced increase in the conservation at position 179. In the carboxyl direction, the amino acids at position 187 are disproportionately conserved. Notably, these flank the CM1 F at position 183 in a helical projection, not the APC D. This is not unusual in active site helices, where the contact of the helix with the bulk of the protein is the most highly conserved, while the active site residues conserved in a portion of the helix 'fade' in one direction as the active site helix moves away from the critical section. Remarkably, at both positions, the predominant amino acid is basic; at position 179 it is Lys, and at position 187 an Arg or Lys. Thus, there is weak evidence for a helical structure here.

The functional subfamilies can also be mapped out on a beta strand template to look for patterns of alternating properties. The underside of the beta strand is clearly more variable than the top side. This is most evident in functional subfamily 6, but also in functional subfamilies 3, and particularly at positions 177–181. Further, the pattern of alternation from positions 177–181 is evident, making the beta structure more plausible for this segment. Thus, a weak case can also be made that this segment adopts a beta structure.

TABLE 2. NUMBER OF CONSERVED SUBGROUPS BY POSITION NUMBER IN CLUSTERS OF SUBGROUPS WITH DIFFERENT MINIMUM PAIRWISE IDENTITIES (MPI)

Position number	MPI (%)						
	90	85	80	70	60	50	40
177	12	4	4	1	0	0	0
178	12	6	5	4	2	1	0
179	12	6	6	6	3	5	1
180	11	5	5	4	1	3	1
181	12	5	5	4	3	1	0
182	12	6	6	6	5	6	4
183	12	6	6	6	5	6	4
184	12	6	6	6	5	6	4
185	12	6	6	6	5	3	1
186	12	6	6	5	5	2	1
187	12	5	5	6	4	4	1
188	12	5	5	3	3	2	0
189	10	5	5	6	4	1	0
190	9	5	1	3	1	1	0
191	10	5	4	3	2	0	0
192	11	4	4	5	0	1	0
193	11	6	4	5	3	0	0
194	9	4	3	6	2	3	0

An expedient that has been used to resolve uncertain cases (Johnsson and Benner, unpublished) has been to reconstruct the sequences of ancient proteins using a rule of parsimony and examine their structures. By this process, weak evidence for a helix can be found in this region; in the tyrosine kinases, positions 177, 181, and 188 are all indeterminate, and all lie on one side of a helical wheel. However, this region in the reconstructed protein is highly hydrophobic. Still stronger evidence for beta strand comes in subgroup 6, where 177, 179, and 185 are indeterminate, 181 nearly so; 183 is an APC F.

This is the most difficult segment of the protein to assign. Crystal structures of several protein kinases must be done to learn how much conformational variation occurs in this region in different subfamilies. The assignment noted in Figure 5 for this region is our best guess, both on the grounds outlined above, and from general intuition.

Unit 21. The deletion at positions 194–197 in functional subfamily 1 indicates that positions 193 and 198 are close in the remaining proteins. This region contains the autophosphorylation site for the tyrosine kinases (position 194), a site that may be near the active site. However, there is no evidence from the alignment that this segment is at the active site. Here again, this is most likely because of substantial functional adaptation that has led to sequence divergence within different functional subfamilies of proteins. Notably, the autophosphorylation site in the serine-threonine kinases (at position 199 in the next segment) does not align with the autophosphorylation site in the tyrosine kinases (position 194) in the alignment prepared by Hanks *et al.* (1).

Unit 22. This segment appears to be another long broken beta strand. The alignment is not well anchored in this segment. Mobashery and Kaiser (28) aligned the TWTLC segment against the NEYTA segment in the tyrosine kinases, based on the fact that the middle residue in both is the autophosphorylation site. The change in the alignment is extremely significant. If the residues from the tyrosine kinases presently matched with positions 198–204 remain, the conformation of the segment would be assigned as an active site coil. If not, it would be assigned as an active site beta strand from position 198–203.

The alignment from position 190–228 is heavily laden with parsing elements. Nine of the 29 positions involve deletion/insertions, and 18 of the 29 positions have P in at least one protein. Thus, a full 24 of the 29 positions (83%) are candidates of varying strengths for parses. There are several strong active site assignments. Position 209 is a CM6 P, canonically assigned as a secondary parse; a PP parsing sequence found in proteins 46, 65, 80, and 84 (MPI < 35%) confirms the assignment of a parse at this position. This parse is adjacent to an APC E, canonically designated an active site residue. A distributed parse is found at positions 203–204. Further, in positions 190–201, every residue is assigned to the surface, although the assignments at positions 200 and 201 are relatively weak. Likewise, the segment from 217–225 is largely assigned to the surface.

The pattern of variation and conservation suggests that this segment almost certainly lies near the active site. From position 202–216 there are no surface assignments, and extremely little variability. Conserved strings and an APC E at position 210 support more strongly an assignment of this segment to the active site. There is ample experimental evidence to confirm this suggestion (29, 30). As noted above, position 194 is the autophosphorylation site of several tyrosine kinases. Short peptides (e.g., LRRASLG) are phosphorylated by the catalytic subunit of bovine cyclic AMP dependent protein kinase. When these are modified to introduce reactive groups, Thr 199 (alignment number) and Cys 201 are modified, and these residues are protected by substrate.

The principal difficulty in assigning conformations to segments such as these is to choose between an assignment as an active site coil, or as a set of short beta strands separated by bulges or turns. The second is favored in this case, as there are a large number of splits in this segment. There is a conserved string in subgroup 2 (MPI = 50%) (VTLWYR), a canonical indicator of an active site string. We have assigned a beta strand from position 201 to position 212, with the strand bent or broken at position 203–204 and again at position 209.

Unit 23. Positions 213–216 are deleted in most of the proteins in the alignment. Further, there is a deletion at position 224, a deletion that possibly can be moved by readjusting the alignment. The issue regarding the intervening segment (positions 217–223) is whether or not it adopts a standard secondary structure. Four consecutive surface assignments (positions 217–220) are canonically designated as a coil. Further, parsing elements are found at positions 218, 219, 220 and 223. A coil is the canonical assignment, with position 221 serving as its hydrophobic anchor, a residue that points inside to hold the otherwise external loop in a defined conformation. There is a string SSS in protein 45 (positions 217–219) which is also suggestive of a coil.

This conformational assignment may not hold for tyrosine kinases (functional subfamilies 8–12), which do not contain parsing strings or deletions in this region, and otherwise do not appear to be as non-structured in this region as the serine-threonine kinases. However, given the problems with the alignment noted above, this matching may be deceptive.

Unit 24. The segment from 226–240 is entirely assigned to the inside; only at position 237 is there an extremely weak surface assignment (two variable subgroups MPI = 50% with only one having a polar residue). This stretch of 15 interior positions is the longest in the protein. The region has a large number of splits at low MPI's. Thus, the segment is canonically assigned as a beta strand. The strand contains some unusual positions. For example, position 230 displays the highest reflexivity in the entire alignment; Ala and Ser are found in many subgroups at various MPI values, but the distribution is not tree-like. This indicates a structural constraint on drift, confirms the assignment as an interior beta strand, and should be the basis for a search for co-varying positions elsewhere in the sequence.

Again, the beta strand is long. It is, however, conveniently bent at positions 232 and 237. Deciding where to end the strand at the carboxyl terminus is problematic, as the deletion at position 241 is placed here solely because of a presumed alignment anchor at position 242. While it is clear that positions 244–245 adopt a coil conformation, it is possible for the beta strand to extend at least as far as position 243.

Unit 25. This segment is assigned a coil conformation because of the distributed parse at positions 244–245, a presumed hydrophobic anchor at position 246, and the conserved G (secondary parse) at position 242. There is no evidence for active site residues in the coil. Positions 246 and 260 are close in space, as indicated by the deletion in protein 46.

Unit 26. This segment is not well anchored, and the single amino acid insertion at position 273 might well be collapsed by a shift of the alignment. This would move the beginning of the parse to position 276, possibly adding 2 more positions to the helix in the previous segment. This revised alignment is used in the analysis here.

There is textbook amphiphilicity from positions 260 to 272, suggesting an assignment of a 10 residue helix in this region (Fig. 4d). Helices of this type are only rarely misassigned, making this one of the strongest predictions in this structure. It then remains to determine how far this helix can be extended in either direction. The amphiphilic pattern is disrupted in the amino direction by the surface assignment at position 262 (which is on the inside face of the helix). Position 262, therefore, may still be a part of the helix; as noted previously, amino acids on the inside ends of amphiphilic helices are often assigned to the surface. However, this implies that the helix does not extend more than a half turn further in the amino direction. This notion is confirmed by the scattered P's at positions 262, 261 and 260. Further, there is a parsing string PXP in positions 260–262 of protein 9, and the parsing string PP in positions 261–262 of protein 90. Based on this reasoning, the assigned helix begins at position 262.

The carboxyl end of the helix is more difficult to assign. Position 273 breaks the amphiphilic pattern of the helix, but the assignment is generated for the alignment with the deletion at position 273. This alignment is modified to remove the deletion in our final analysis. Likewise, the 'correct' (presuming the helical wheel) assignment of position 274 to the surface is less strong in the realigned segment. To match parses (often in surface helices, the parses at the beginning and ends of the helix fall on the same side of a helical projection), there are several choices, as the potential parses at the beginning of the helix at positions 61, 62, and 63 each match with potential parses at the end (261 matches with 272; 262 matches with 273, and 263 matches with 274 in the rearranged alignment). Nevertheless, it is clear that the helix does not extend past position 275 (note the PPP parsing string in protein 42). Thus, positions 262–273 are assigned a helical structure, with a coil extending from 274 until the next parse.

Buechler and Taylor (Reference 23) found that the carboxyl side chain of D 260 is especially reactive with water-soluble carbodiimide. However, the side chain is not protected by substrate.

Unit 27. Positions 276–285 are deleted in protein 46, indicating that positions 275 and 286 are close in space in the folded structure in the remaining proteins. A large number of parsing strings within the region make the canonical assignment a coil for this segment. There is no evidence for active site residues in the coil.

Units 28, 29 and 30. There is textbook amphiphilicity from positions 288 to 300, suggesting a 15 residue helix (a bit over four turns) in this region (Fig. 4e). Helices of this type and length are almost never misassigned, making this a 'very strong' assignment. It remains to determine how far this helix can be extended in either direction. A deletion parse ending at position 285 suggests that the helix begins exactly at position 286. However, some of the amino acids can be moved from above the deletion to position 285 without significantly altering the significance of the alignment in this region. Indeed, in some proteins (e.g., proteins 52–54), the pattern of amphiphilicity appears to extend in this direction. Extending the helix back by one amino acid also superimposes the parses at each end of the helix. On these grounds, the amino end of the helix is extended to position 284 in the maximum helix in Figure 5 in a readjusted alignment (some residues from positions 279 and 280 of proteins in functional subfamilies 1 and 2 are moved to position 284 in the alignment).

At the carboxyl end of the helix, amphiphilicity is broken at position 301. Further, a secondary parse at position 302 provides a plausible point to end the helix. Positions 299 to 304 are also a string of 6 consecutive surface assignments. Thus, positions 286–301 are assigned a helical structure in the preferred assignment for cAMP-dependent protein kinases (Fig. 5).

The conformations of the subsegments that immediately follow (positions 303–313) are difficult to assign. Positions 303, 304, 307, 309, 310, 313, and 314 are all assigned to the surface. Positions 306, 308, 311, and 312 are assigned to the inside. The most distinctive feature of this segment is the APC R at position 305. This is canonically assigned to the active site (although again, such assignments are only ca. 70% accurate).

Secondary parses at positions 302 and 306 separated by 2 surface residues are canonically assigned as a loop. There are no active site strings that would confirm the APC R as an active site residue, even at relatively high MPI values. Thus, it might be argued that the APC R is an 'anchor' for a loop not at the active site. However, in view of the wide range of functions performed by different members of this family of proteins, it is important to reexamine this point by functional class, assuming that the variation that is seen is adaptive in a substrate binding segment of the kinases. For example, in functional group 1 (50% MPI), one does not find the amount of variation that one would expect in a surface loop far from the active site. Thus, this is assigned as an active site coil.

The conformation of the following segment is especially problematic. Canonically, the presence of parsing strings (e.g., in functional subfamily 2, GSGPDGEP) and weak secondary parsing elements at positions 307, 309, 310, and 312 suggests that this segment is a coil with hydrophobic anchors at positions 308 and (in some subfamilies) 311 and 312. However, position 307 shows an intriguing amount of reflexivity. For example, in functional subfamily 6, with a MPI of ca. 30%, only residues T and S are found, and this variation is mirrored in functional subfamilies 3, 4, 5, 9 and 11. Such reflexivity suggests the possibility that this segment is a beta strand, possibly near the surface.

Unit 31. Positions 314–318 are deleted in functional subgroups 3-6, implying that positions 313 and 319 are close together in space in the remaining proteins. There is no evidence for active site residues in the coil.

Unit 32. It is important to note that the sequences in this alignment are truncated; they continue past this point, often for some length. However, these carboxyl terminal extensions in the different proteins cannot be aligned. Canonically, it is simplest to assign this segment simply as a coil, as it lies largely on the surface (all positions are assigned to the surface). However, the strength of the assignments vary as expected for an alpha helix. So does the hydrophobicity of the segment. Thus, it is possible that this segment is the beginning of a helix that will extend into the next section of the protein. It is important to test this by looking at the extended sequences of proteins, where they can be aligned.

In functional group 1, 2 and 8, the amphiphilic helix can be extended by one position (to 326). Group 8 has a secondary parse at what would be position 327 in the alignment. In the serine kinase groups, this position is occupied by a conserved W, a W that breaks the amphiphilic helix at this position. The sequence that follows also does not fit on the amphiphilic wheel; the next parsing segment in this class is at positions 337. While the method cannot discuss these structures without an alignment, there is no reason not to accept in this segment a short helix (7 amino acids, 319–325, with 319 being a weak surface). The P scattered in functional groups 1–5 either indicate that the helix is shorter in these groups (in ADH, phospholipase, and other proteins there are analogous single turn helices that show amphiphilicity) or, more probably in our opinion, a P occurs in the first turn of an alpha helix.

Assembling the Secondary Structural Elements

To assemble the secondary structural units, positions 46, 67, 113, 116, 156–162, 182, 199, 201, 210, 237, and 305 are tentatively assigned as lying at or near the active site, suggesting that these points in the polypeptide chain come together in 3-dimensional space. Position 305 is only weakly assigned to the active site, however, and alternative models that place this residue at a position removed from the active site must also be considered.

The beta strands that are connected by short loops are then examined to see if evidence can be found that they lie antiparallel in a beta sheet. As discussed above, co-variation analysis suggests that the two strands 84–93 and 103–111 lie antiparallel, with the side chain of the amino acid at position 87 of the first beta strand on the same side of the sheet as the side chain of the amino acid at position 108, and these two side chains in close proximity.

Further, based on active site assignments, beta strand 201–212 is arranged antiparallel to beta strand 226–240, although these long strands are almost certainly discontinuous, implying that this antiparallel arrangement need not extend along the entire strand.

Chemical cross-linking experiments suggest that the side chain of the amino acid at position 46 in the beta strand 43–48 is in close proximity to the side chain at position 182 in the beta strand 177–185; co-variation analysis suggests that these these strands lie antiparallel (based on functional variation at positions 47 and 181, and the hydrophobic variable position 85 and the APC F at position 198). In contrast, no evidence can be found for any association of beta strand 12–22 with beta strand 43–48.

These considerations lead to the minimal structure shown in Figure 6. It should be noted that in this figure the long beta strands are almost certainly bent. Further, constraints imposed by a need for a 2-dimensional representation distort the picture. Thus, as indicated on the figure, the structure is further folded to bring together the active site residue at position 67 (near the ATP binding site) and at 113–116 (near the peptide binding site).

This picture can be modified by inferences drawn from chemical considerations of the reaction being catalyzed. First, the nucleophilic displacement is 'in-line' with a trigonal bipyramidal phosphorus in the transition state lying between an attacking nucleophile and the departing beta-phosphate of ATP. The nucleophile (a serine, threonine or tyrosine) must lose a proton to a basic residue at the active site. The pentaco-ordinate phosphate with additional negative charge must be stabilized by a positively charged residue on the protein. The Mg–ATP must be bound on the distal side of the phosphorus, with the divalent magnesium cation co-ordinated to the alpha and beta phosphates of ATP and to ligands on the enzyme. Finally, groups on the enzyme must form hydrogen bonds to the ribose ring hydroxyl groups, and present a hydrophobic pocket to hold the purine ring with a hydrogen bond to N(6).

What sorts of residues might one expect the enzyme to contribute for this sort of catalytic effect? Some information bearing on this question can be obtained from the crystal structures of phosphoglycerate kinase (31), phosphofructokinase (32), adenylate kinase (33) and pyruvate kinase (34).

Consider first the co-ordination of magnesium. In pyruvate kinase, the magnesium co-ordinates to the side chain of Glu 271 and two main chain carbonyl residues. Glu 271 is conserved in enzymes from cat, chick, rat, and yeast, is part of a conserved string of three, and lies in a turn at the end of a beta strand and two positions before the start of an alpha helix. In adenylate kinase, the side chain of Asp 93 (in the middle of a beta strand) appears to co-ordinate the magnesium. In phosphofructokinase, Asp 103 (in a conserved heptapeptide at the start of an alpha helix) co-ordinates magnesium. Asp 129 appears to bind water co-ordinated to magnesium in this protein. In phosphoglycerate kinase, Asp 374 (at the start of an alpha helix in a GGGD conserved string) co-ordinates the magnesium. Thus, we can expect that Asp and Glu residues (with perhaps a slight preference for the former) will also be involved in the co-ordination of magnesium in protein kinase, and that these residues can be identified by their pattern of conservation.

Interactions with the phosphate groups involve, not surprisingly, Arg and Lys residues, with perhaps a preference for the former. For example, in pyruvate kinase, Lys 114 and Arg 119 form salt bridges to the phosphate groups of the substrates. Arg 72 (in a beta strand) and Arg 293 (in a coil between a beta strand and an alpha helix) appear to be near the phosphorus electrophile. Both are highly conserved, the second in a string many amino acids long. In adenylate kinase, Lys 21 (in a turn or at the beginning of a helix) may be near the gamma-phosphate group. The residue is conserved in five sequences with pairwise identities from 24% to 52%. Arg 44, Arg 97, Arg 128, Arg 138, and Arg 149 are all conserved, and all point towards the active site cleft in this enzyme, not surprising considering the number of negatively charged groups that are brought together in the active site. In phosphofructokinase, Arg 72 seems to bind to the alpha-phosphorus of ATP as well as the phosphorus being transferred. In phosphoglycerate kinase, Lys 219 (at the start of an alpha helix) appears to bind the alpha-phosphate of ATP. The beta and gamma phosphates are found on the amino terminal end of a helix.

Co-ordination to ribose is varied. In phosphofructokinase, Tyr 41 (in a conserved dipeptide at the start of an alpha helix) provides the contact. In phosphoglycerate kinase, Glu 343 (in a coil at the end of a GVFE conserved sequence) appears to make a hydrogen bond with the ribose hydroxyl groups. In alcohol dehydrogenases, the ribose of NAD$^+$ is hydrogen bonded to an APC Asp 223 (at the end of a beta strand).

It is now necessary to assign similar roles to various positions in protein kinase. In addition to the sequences themselves, we have some information available to us from chemical modification studies involving carbodiimide (for acidic residues), acetic anhydride (for Lys), and peptide analogs, obtained primarily from the laboratories of Taylor and Kaiser, which are summarized in Table 3.

There are only two candidates for a side chain to provide the positive charge to the gamma phosphate, K 46 and R 305. The chemical modification experiments with reactive analogs of ATP place the side chain K 46 near the gamma-phosphorus of ATP, and it seems not unreasonable to assign tentatively this role to this residue.

Identifying roles for the individual carboxyl groups is more difficult, as they can presumably serve to co-ordinate magnesium, or bind to the ribose, or act as a base abstracting a proton from the nucleophilic center of the protein substrate. Acidic residues under consideration here are of three types. First, there are two residues that are conserved, reactive with carbodiimide, and are protected by Mg–ATP (E 067 and D 182). Second, there are two residues that are conserved and do not react with carbodiimide (D 157 and E 210). Then there are two acidic residues that are not conserved, react with carbodiimide, but require both Mg–ATP and peptide to protect them (E 161 and E 237). Finally, there are two acidic residues that are not conserved, reactive, and are not protected by any sort of substrate combination (E 089 and D 260).

TABLE 3. SUMMARY OF CHEMICAL MODIFICATION DATA AS IT PERTAINS TO ASSIGNING POSITIONS TO THE ACTIVE SITE

Position	Relevant protection	Conservation	Assigned role
K 007	MgATP + peptide	Not conserved	Binds peptide
K 046	MgATP	Conserved	Gamma P
K 050	MgATP + peptide	Not conserved	Binds peptide
E 067	MgATP	Conserved	Mg ligand
E 089	Not protected	Not conserved	No role
G 113	Peptide	Not conserved	Binds peptide
M 116	Peptide	Not conserved	Binds peptide
D 157	Not labeled	Conserved	Mg ligand
N 162	No information	Conserved	Binds ribose
E 161	MgATP + peptide	Not conserved	AS vicinity
D 182	MgATP	Conserved	General base
T 199	Peptide	Not conserved	Binds peptide
C 201	Peptide	Not conserved	Binds peptide
E 210	Not labeled	Conserved	Structural
D 227	Not labeled	Almost conserved	Structural
E 237	MgATP + peptide	Not conserved	AS no role
D 260	Not protected	Not conserved	No role
R 305	No information	Conserved	Uncertain

The last pair, E 089 and D 260, can be presumed not to lie at the active site, and therefore to be uninteresting for the present discussion. E 161 and E 237 might be presumed to lie between the substrate peptide binding site and the Mg–ATP binding site. Because they are not conserved, they presumably play no role in catalysis. However, these positions in the polypeptide chain should lie between positions 113–116 and 199–201 (the peptide binding site) and positions 157–160 and 67 (the Mg–ATP binding site).

In contrast, both D 157 and E 210 are absolutely conserved. Both do not react with carbodiimide, suggesting either that they are buried, or that they are for chemical reasons unreactive. The assignment of E 210 to the active site is not extremely strong. It is in a beta strand that appears to be inside the folded structure, but displays few of the sequence characteristics expected for a string at the active site. As noted above, absolutely conserved charged residues appear only ca. 70% of the time at the active site. It is possible that E 210 is one of the 30% of this type of position that does not.

A possibly analogous residue is D 227. It is nearly, but not completely, conserved, being substituted by Ala in protein 46. Such variation is not expected for a residue critically involved in catalysis, although it is worth noting that the variation could be a sequencing error. Further, D 227 occurs again in a buried beta strand, and again might be highly conserved because it plays a structural, rather than a catalytic role.

In contrast, there is no question from the pattern of conservation in the region that D 157 lies at the active site. It must therefore be accessible to solvent under some conditions. We cannot say why it fails to react with carbodiimide. However, it is worth noting that the reaction with carbodiimide proceeds from the protonated form of a carboxyl group. A carboxyl group with a low pKa might not be expected to react. This is a behavior exactly opposite that expected for a carboxylate involved as a general base catalyst, suggesting that the residue at position 157 does not play this role. Presumably E 067 acts either as a ligand for magnesium or to form hydrogen bonds with the hydroxyl groups on the ribose ring.

The conservation of E 067 and D 182 indicates that these residues do play critical roles. However, in contrast to D 157, D 182 is extremely reactive with carbodiimide, suggesting that the side chain carboxylic acid has a high pKa. This implies, of course, that the conjugate base is relatively strong, as would be expected were this carboxylate to act as a base to deprotonate the nucleophile in the reaction. In this role, D 182 should be close enough to the position where the gamma-phosphate would bind in the enzyme–substrate complex that it forms a cross-link with K 46 when no substrate is present.

In this context, it is worth noting that the pKa values of the hydroxyl groups of serine and tyrosine are different by some six orders of magnitude. Thus, it is not expected *a priori* that the same basic group will abstract a proton from each. We had hoped to find in the protein kinase alignment a basic residue conserved in each functional subfamily but different between the two (a 'functional split'), and use this pattern to identify the residue that acts as a base. Unfortunately, no such residue appears to exist in the alignment. Closest are positions 149 (mostly His in the serine/threonine kinases, and mostly Glu in the tyrosine kinases), but the conservation in each functional subfamily is not perfect. Thus, our preferred candidate for a residue that acts as a general base in the catalytic reaction is Asp 182.

These conclusions can be summarized in a model of the active site (Fig. 7) showing the transition state for the reaction and the presumed orientation of reactions involved in catalysis. While both E 67 and D 157 are shown co-ordinating magnesium, one of these two side chains might instead be forming hydrogen bonds to the hydroxyl groups of the ribose ring.

The models shown in Figures 6 and 7 are not, of course, complete representations of the conformation of the folded protein. However, provided that the two peptide binding sites are brought together in three dimensions, positions 67 and 157 brought to within bonding distance of a single magnesium, positions 161 and 237 placed between the peptide binding site and the ATP binding site, and position 46 aligned with the gamma phosphate and 182 placed to abstract a proton from the reactive nucleophile, only a small number of folded forms are consistent with the constraints imposed by the assignments of secondary structure, the assignments of inside and surface positions (Fig. 3), and the covalent links of the polypeptide backbone. Further constraints are introduced if position 7 is placed near the active site.

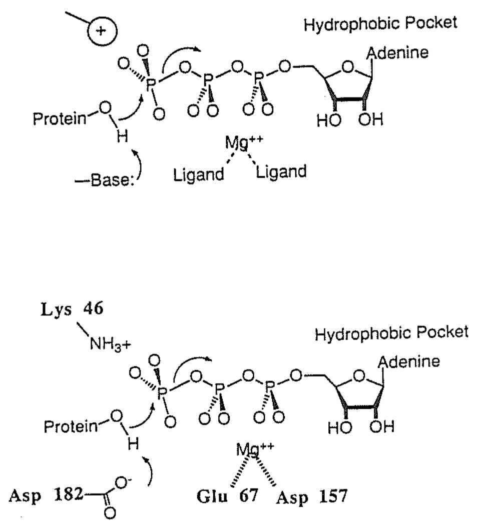


FIG. 7. Diagram of the active site of protein kinase showing transition state for the reaction catalyzed and the amino acids presumed to be involved in catalysis.

We do not have sufficient computer expertise at this point to go the last step and incorporate all of these constraints into a 3-dimensional model. Especially important in this regard is to bend the beta strands correctly (at the positions indicated by the dotted lines), to adjust the twist of the appropriate beta sheets, to order the beta strands based on their general accessibility (for example, beta strand 226–241 is undoubtedly in the center of a beta sheet, not on the end as a naive reading of Figure 6 would imply), and to pack interior residues into a plausible core. Work in these areas is in progress.

SUMMARY

The secondary structure and elements of tertiary structure have been predicted for the catalytic domain of protein kinases using a method that extracts structural information from the patterns of conservation and variation in an alignment of homologous proteins (35). The central features of this structural prediction are: (a) the catalytic domains of protein kinases do not incorporate a Rossmann fold; (b) the core of the structure is founded on beta sheets built from pairs of bent antiparallel beta strands; (c) five helices, including an especially long helix (alignment positions 129–152) that lie on the outside of the folded core. These proteins are important in many aspects of metabolic regulation.

ACKNOWLEDGEMENTS

We are indebted to the Swiss National Science Foundation and Sandoz for support of this work.

REFERENCES

1. S. K. HANKS, A. M. QUINN and T. HUNTER, The protein kinase family: Conserved features and deduced phylogeny of the catalytic domains, *Science* **241**, 42–52 (1988).
2. Y. NISHIZUKA, The molecular heterogeneity of protein kinase C and its implications for cellular regulation, *Nature* **334**, 661–665 (1988).
3. S. S. TAYLOR, cAMP-dependent protein kinase: Model for an enzyme family, *J. Biol. Chem.* **264**, 8443–8446 (1989).
4. S. A. BENNER, Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure, *Advan. Enzyme Regul.* **28**, 219–236 (1989).
5. S. A. BENNER and A. D. ELLINGTON, Evolution and structural theory: The frontier between chemistry and biochemistry, *Bioorg. Chem. Frontiers* **1**, 1–70 (1990).
6. M. J. E. STERNBERG, Deciphering the protein folding code, *Trends Biochem. Sci.* **15**, 360–361 (1990).
7. G. FASMAN, (ed.), *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum New York (1989).
8. C. CHOTHIA and A. M. LESK, The relationship between the divergence of sequence and structure in proteins, *EMBO J.* **5**, 823–826 (1986).
9. L. PATTHY, Prediction of surface loops of protein-folds from multiple alignments of homologous sequences, *Acta Biochim. Biophys. Hung.* **24**, 3–13 (1989).
10. F. E. COHEN, R. M. ABARBANEL, I. D. KUNTZ and R. J. FLETTERICK, Turn prediction in proteins using a pattern-matching approach, *Biochemistry* **25**, 266–275 (1986).
11. J. S. RICHARDSON, Anatomy and taxonomy of proteins, *Advan. Prot. Chem.* **34**, 167–339 (1981).
12. J. M. SOWADSKI, N. H. XUONG, D. A. ANDERSON and S. S. TAYLOR, Crystallization of cAMP-dependent protein kinase. Crystals of catalytic subunit diffract to 3.5 Å resolution, *J. Mol. Biol.* **182**, 617–620 (1985).
13. S. SHOJI, D. C. PARMELEE, R. D. WADE, S. KUMAR, L. H. ERICSSON, K. A. WALSH, H. NEURATH, G. L. LONG, J. G. DEMAILLE, E. H. FISCHER and K. TITANI, Complete amino acid sequence of the catalytic subunit of bovine cardiac muscle cyclic AMP-dependent protein kinase, *Proc. Natl. Acad. Sci.* **78**, 848–851 (1981).
14. I. P. CRAWFORD, T. NIERMANN and K. KIRSCHNER, Prediction of evolutionary structure by evolutionary comparison: Application to the alpha subunit of tryptophan synthase, *Proteins* **2**, 118–129 (1987).
15. S. R. ADAVANI, M. SCHWARZ, M. O. SHOWERS, R. A. MAURER and B. A. HEMMINGS, Multiple RNA species code for the catalytic subunit of cAMP-dependent protein kinase from LLC-PK 1 cells, *Eur. J. Biochem.* **167**, 221–226 (1987).
16. E. A. FIRST and S. S. TAYLOR, Selective modification of the catalytic subunit of cAMP-dependent protein kinase with sulfhydryl-specific fluorescent probes, *Biochemistry* **28**, 3598–3605 (1989).
17. F. E. COHEN, R. M. ABARBANEL, I. D. KUNTZ, and R. J. FLETTERICK, Secondary structure assignment for alpha/beta proteins by a combinatorial approach, *Biochemistry* **22**, 4894–4904 (1983).

18. M. J. E. STERNBERG and J. E. THORNTON, On the conformation of proteins: An analysis of beta-pleated sheets, *J. Mol. Biol.* **110**, 285–296 (1977).
19. R. K. WIERENGA and W. G. J. HOL, Predicted nucleotide-binding properties of p21 protein and its cancer-associated variant, *Nature* **302**, 842–844 (1983).
20. M. J. E. STERNBERG and W. R. TAYLOR, Modelling the ATP-binding site of oncogene products, the epidermal growth factor receptor and related proteins, *FEBS Lett.* **175**, 387–392 (1984).
21. J. A. BUECHLER, T. A. VEDVICK and S. S. TAYLOR, Differential labeling of the catalytic subunit of cAMP-dependent protein kinase with acetic anhydride: Substrate-induced conformational changes, *Biochemistry* **28**, 3018–3024 (1989).
22. M. J. ZOLLER, N. C. NELSON and S. S. TAYLOR, Affinity labeling of cAMP-dependent protein kinase with *p*-fluorosulfonylbenzoyladenosine, *J. Biol. Chem.* **256**, 10837–10842 (1981).
23. J. A. BUECHLER and S. S. TAYLOR, Identification of aspartate-184 as an essential residue in the catalytic subunit of cAMP-dependent protein kinase, *Biochemistry* **27**, 7356–7361 (1988).
24. J. A. BUECHLER and S. S. TAYLOR, Differential labeling of the catalytic subunit of cAMP-dependent protein kinase with a water-soluble carbodiimide: Identification of carboxyl groups protected by MgATP and inhibitor peptides, *Biochemistry* **29**, 1937–1943 (1990).
25. W. T. MILLER and E. T. KAISER, Probing the peptide binding site of the cAMP-dependent protein kinase by using a peptide-based photoaffinity label, *Proc. Natl. Acad. Sci.* **85**, 5429–5433 (1988).
26. J. A. BUECHLER and S. S. TAYLOR, Identification of the peptide recognition site in the catalytic subunit of cAMP-dependent protein kinase, *J. Cell Biol.* **107**, 491a-(1989).
27. J. A. BUECHLER and S. S. TAYLOR, Dicyclohexylcarbodiimide cross-links two conserved residues, Asp-184 and Lys-72, at the active site of the catalytic subunit of cAMP-dependent protein kinase, *Biochemistry* **28**, 2065–2070 (1989).
28. S. MOBASHERY and E. T. KAISER, Identification of amino acid residues involved in substrate recognition by the catalytic subunit of bovine cyclic AMP dependent protein kinase: Peptide-based affinity labels, *Biochemistry* **27**, 3691–3696 (1988).
29 H. N. BRAMSON, N. THOMAS, R. MATSUEDA, N. C. NELSON, S. S. TAYLOR and E. T. KAISER, Modification of the catalytic subunit of bovine heart cAMP dependent protein kinase with affinity labels related to peptide substrates, *J. Biol. Chem.* **257**, 10575–10581 (1982).
30. N. C. NELSON and S. S. TAYLOR, Differential labeling and identification of the cysteine-containing tryptic peptides of the catalytic subunit of porcine heart cAMP-dependent protein kinase, *J. Biol. Chem.* **256**, 3743–3750 (1981).
31. R. D. BANKS, C. C. F. BLAKE, P. R. EVANS, R. HASER, D. W. RICE, G. W. HARDY, M. MERRETT and A. W. PHILLIPS, Sequence, structure, and activity of phosphoglycerate kinase: a possible hinge-bending enzyme, *Nature* **279**, 773–777 (1979).
32. Y. SHIRAKIHARA and P. R. EVANS, Crystal structure of the complex of phosphofructokinase from *Escherichia coli* with its reaction products, *J. Mol. Biol.* **204**, 973–994 (1988).
33. D. DREUSICKE, P. A. KARPLUS and G. E. SCHULZ, Refined structure of porcine cytosolic adenylate kinase at 2.1 Å resolution, *J. Mol. Biol.* **199**, 359–371 (1988).
34. H. MUIRHEAD, D. A. CLAYDEN, D. BARFORD, C. G. LORIMER, L. A. FOTHERGILL-GILMORE, E. SCHLITZ and W. SCHMITT, The structure of cat muscle pyruvate kinase, *EMBO J.* **5**, 475–481 (1986).
35. J. REED and V. KINZEL, Near- and far-ultraviolet circular dichroism of the catalytic subunit of adenosine cyclic 5′-monophosphate dependent protein kinase, *Biochemistry* **23**, 1357–1362 (1984).