Predicting Phosphorylation: A critique of the NetPhos program and potential alternatives

Andy Poon

Biochemistry 218-Spring 2004

Introduction and Purpose

Posttranslational modification of proteins generates dynamic mechanisms by which proteins can switch from one state to another. Such modifications are integral to the modulation of various cellular functions, from directing cellular localization to regulating protein activity. Perhaps the most prevalent of these modifications is the phosphorylation and dephosphorylation of a substrate by kinases and phosphatases, respectively.
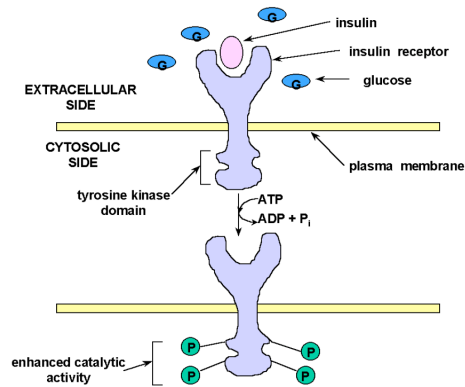
Phosphorylation is an essential method through which various signaling pathways operate. The addition or removal of a phosphate group from an amino acid side chain may cause conformational changes throughout the protein structure which affect its function or localization. The presence of a phosphate group often modulates that protein's interaction with other proteins upstream or downstream of it on a certain pathway, such as the binding of phosphorylated receptors by proteins with SH2 domains (reviewed by Pawson, 2004).

Given the prevalence of phosphorylation as a regulatory mechanism in cellular functions, there exists a need to predict whether certain proteins are susceptible to this posttranslational modification. When searching through an amino acid sequence for

potential phosphorylation sites, the only certainty is that the phosphorylated residue must be a serine, threonine, or tyrosine. These three amino acids have the capacity to bind phosphates because they all contain a hydroxyl group in their side chains which are deprotonated at physiological pH, such that the oxyanion can act as a nucleophile to attack a phosphate from ATP. However, not all serines (S), threonines (T), and tyrosines (Y) are susceptible to phosphorylation.

The likelihood of a given S/T/Y residue to be phosphorylated is defined by other sequences in the protein. Residues close to the S/T/Y in question may define this context, such that certain acidic or basic side chains are preferentially placed next to the phosphorylated residue in order to facilitate the phosphorylation reaction. In addition, this context may be affected by sequences or motifs in the protein which are far away from the putatively altered residue. There are two possible explanations for how residues far away in sequence space from the putative site can dictate phosphorylation. One possibility is that these residues, though far away in sequence space, may be close to the putative site in tertiary structure of the protein, such that it assists in coordinating the docking of a kinase to the protein. A second, more general possibility is that the presence (or absence) of certain residues at a particular position of a protein can dictate indirect allosteric effects which promote phosphorylation. For example, insulin binds to certain motifs on insulin receptors, causing an allosteric change in the receptor structure, allowing for the insulin receptor to be autophosphorylated. Therefore, in this example, motifs allotted for small molecule (i.e. insulin) binding may be indicative of phosphoproteins and thus prove useful in predicting proteins which undergo phosphorylation (Figure 1).

Figure 1: Insulin receptor autophosphorylation. Insulin recognizes a certain motif on the extracellular surface of the receptor, causing autophosphorylation. This motif may be a general indicator for phosphoproteins. **(**adapted from Stryer, Biochemistry; 996-97, 1988)

Currently, there exists one program which seeks to predict phosphorylation sites from a sequence input. NetPhos, available at http://www.cbs.dtu.dk/services/NetPhos/, is a program developed in 1999 which predicts sites by analyzing the sequence space of around any serine, threonine, and tyrosine found in a given sequence (Blom et al, 1999). Positive results are based largely around comparisons of sequence and structure to verified phosphorylation sites from the PhosphoBase database (http://www.cbs.dtu.dk/databases/PhosphoBase) developed by the same research group.
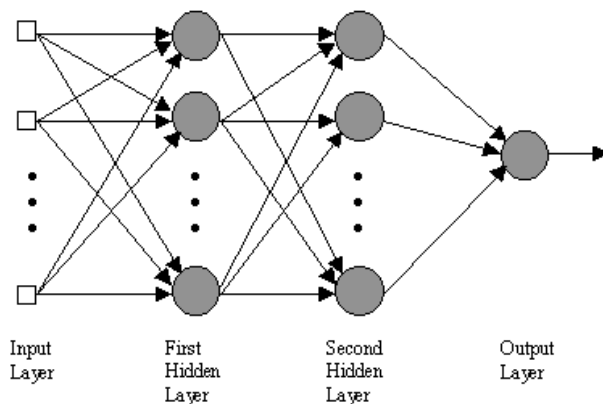
This research paper seeks to critically evaluate the ability of the NetPhos program to predict true phosphorylation sites from an amino acid sequence, assessing the underlying assumptions under which the program functions. Given the lack of bioinformatics tools available to analyze this phenomenon, the latter portion of this paper explores alternative methods for predicting phosphorylation sites. Primarily, the search for alternatives will involve using various methods for pattern matching and sequence similarity searches with phosphorylated proteins.

The NetPhos approach

The NetPhos program utilizes an artificial neural network to predict

phosphorylation sites from an inputted amino acid sequence (Blom et al, 1999). An

artificial neural network is designed upon the principles of how a true biological neural

network operates, where a given node may have tens of thousands of inputs and a similar

number of outputs (Figure 2). Artificial neural networks have been used in a variety of

prediction algorithms, including other analyses of posttranslational modification such as

glycosylation (reviewed by Nakai, 2001).

Figure 2: Organization of an artificial neural network.

(source: http://www.nd.com/neurosolutions/products/ns/whatisNN.html)

The strength of a neural network comes from its ability to adapt to circumstances

to recognize trends from complex and imprecise data. Neural networks are configured to

an application through extensive training by exposing it to various samples. For the

NetPhos program, its neural network is designed for pattern recognition. The creators of

this artificial neural network introduced NetPhos to the PhosphoBase database of

phosphorylated proteins for training. The data in PhosphoBase consists entirely of

sequences. For any phosphorylated protein, the database provides the entire sequence of

414 phosphoproteins followed by remarks on which residues are phosphorylated within the context of +/- 4 residues flanking the phosphate acceptor (Kreegipuu et al, 1999). Therefore, the NetPhos artifical neural network has acquired its ability to identify novel phosphorylation sites by identifying common prevalent features of the sequences immediately surrounding the phosphoacceptor residue (S/T/Y) of known phosphoproteins.

The ability of any artificial neural network to succeed in interpreting new data depends on the quality and quantity of the existing data from which the network was trained. In the case of NetPhos, this artificial neural network was trained from less than 500 proteins. The number of phosphoproteins identified since 1999, when the most current version of PhosphoBase was gathered, has certainly increased several folds. Given that a given phosphoprotein is likely to have more than one phosphorylated residue, the amount of current, relevant information missing from the NetPhos program may cause the performance of the program to be sub-optimal in predicting phosphorylation sites. According to the PhosphoBase developer, as high as 50% of all proteins may be phosphorylated (Kreegipuu et al, 1999). As artificial neural networks are not only developed from examples, but limited by them, the NetPhos program needs to be constantly updated for it to be an effective predictor. The NetPhos neural network has not been retrained since its inception in 1999.

The NetPhos program acquired its sequence recognition capabilities by analyzing the sequence context within which the phosphorylated residue lies. The programmers of NetPhos underwent an exhaustive study to determine whether certain amino acids were preferred at certain positions surrounding the phosphorylated residue (Blom et al, 1999).

For all the verified phosphorylated serines, threonines, and tyrosines from PhosphoBase, the twelve amino acids before and after the acceptor residue were recorded. The researchers uncovered some statistically significant trends. For example, 78% of phosphorylated tyrosines have at least one acidic residue within the four residues before the tyrosine. For phosphorylated serines and threonines, basic residues tend to lay +/- 3 residues from the phosphorylated site. Also, a proline is vastly over-represented at the site immediately C-terminal (position +1) to the phosphorylated residue. Overall, the researchers noted that certain residues were more frequently found around a phosphorylated S/T/Y than normal.

The developers of the program operate under the assumption that "determinants of phosphorylation sites probably are no longer than about ten residues" (Blom et al, 1999). While this position seems moderately justified, given the trends in amino acid positioning surrounding the phosphorylation site stated above, this does not account for the possibility that other determinants of phosphorylation exist outside of the sequences immediately adjacent to the acceptor residue. When a kinase or phosphatase encounters a potential substrate for phosphorylation or dephosphorylation, it recognizes the three-dimensional structure of the protein. Given the various torsion angles which polypeptides may adopt when folded, the residues which are spatially closest to the phosphorylated residue are unlikely to be those amino acids closest in sequence space, but possibly tens or hundreds of amino acids downstream or upstream.

The developers of the NetPhos program also attempted to address the spatial context of phosphorylation sites by creating a different neural network based on structure. This network was generated using distance matrices acquired from PDB protein

structures. Unfortunately, this particular algorithm produced a high level of false positives (29%-41%). This may not be surprising, given that the data set from which this neural network was nurtured must surely be even smaller than the 414 from PhosphoBase, since it is unlikely that protein structures exist for all of the verified phosphoproteins. As a result, current NetPhos program only uses the sequence-based prediction algorithm, and not this structure-based neural network.

After an amino acid sequence is inputted into the NetPhos interface, the program outputs a text page which includes a score ranging from 0-1 for each serine, threonine, and tyrosine in the sequence (Figure 3). For every residue with a NetPhos score higher than 0.5, the program identifies it as a potential hit. Of course, the NetPhos score of 0.5, which the program designates as the minimum score for potential phosphorylation sites, is entirely arbitrary. Therefore, the amount of false positives can automatically be reduced by raising this arbitrary threshold, but at the expense of eliminating some true positives.

Figure 3: Sample NetPhos output (for human Wee1).


```
Phosphorylation sites predicted:      Ser: 36  Thr: 12  Tyr: 3


               Serine predictions

Name           Pos   Context    Score  Pred
_____v_____
gi_1351419_      3    --XMSFLSR  0.002    .
gi_1351419_      6    MSFLSRQQP  0.050    .
gi_1351419_     29    KLIFSPCSD  0.476    .
gi_1351419_     32    FSPCSDCEE  0.979   *S*
gi_1351419_     46    EEEGSGHST  0.751   *S*
gi_1351419_     49    GSGHSTGED  0.992   *S*
gi_1351419_     54    TGEDSAFQE  0.538   *S*
gi_1351419_     61    QEPDSPLPP  0.977   *S*
gi_1351419_     68    PPARSPTEP  0.987   *S*
gi_1351419_     79    ERRRSPGPA  0.998   *S*
```

The most limiting aspect of the NetPhos program is the large number of false positives one receives from the output. The program vastly overestimates the number of residues that are phosphorylated in a given protein sequence. For example, when the

sequence for the human cell cycle protein Wee1 (ascension number: P30291) was entered

into the NetPhos program, the program identified approximately 50% (51 out of 107) of

all serines, threonines and tyrosines in the sequence as potentially phosphorylated.

Despite the prevalence of phosphorylation as a crucial modification in controlling cellular

events, particularly during the cell cycle, it is highly unlikely that a given phosphoprotein

can be phosphorylated at over 50 residues. Recent work identifying the sites of

phosphorylation on Xenopus Wee1 using mass spectrometry has shown that it is

phosphorylated at 11 sites (Kim, SY and Ferrell, JE, personal communication).

Consequently, if one attempts to identify phosphorylation sites of Wee1, the NetPhos

program would only succeeded in eliminating one-half of the S/T/Y residues as being

phosphorylated, leaving over 50 residues remaining to be experimentally verified as true

or false sites, of which approximately 80% were false.

Overall, the NetPhos program succeeds in narrowing down the possibilities for

phosphorylation sites for a given phosphoprotein, particularly since it is currently the

only available bioinformatics resource for that purpose. However, the program does a

poor job of distinguishing whether an unknown protein is phosphorylated at all or not.

Due to high number of true positives, NetPhos would likely identify virtually any

sequence with a serine, threonine, or tyrosine as a potential phosphoprotein.   Considering

the fact that NetPhos also fails to consider three-dimensional positioning in its algorithm,

the need exists for alternative methods of determining phosphorylation state of proteins.

Alternative methods for phospho-prediction

The NetPhos program recognizes patterns in sequences in close proximity to the

phosphorylated residue. For example, NetPhos may recognize that an arginine located

two amino acids before the potential acceptor residue is indicative of phosphorylation. A reasonable alternative selection criterion may involve recognizing whole motifs rather than just the relative location of certain residues relative to the phospho-serine/threonine/tyrosine. Motifs indicative of phosphorylation may include sequences which are close to the phosphorylated residue. For example, RX*(S/T)*P is a commonly accepted motif for phosphorylation by cAMP-dependent protein kinase (with the bold-italicized "S/T" indicative of the phosphorylated serine or threonine). However, motifs indicative of phosphorylation may also exist much further upstream or downstream of a putative phosphorylated residue. These "far-away" motifs may influence phosphorylation by perhaps being docking sites for kinases. Or perhaps these "far-away" motifs may actually be in close 3-D spatial proximity to the phosphorylated residue despite being distant in sequence.

To examine the possibility that certain motifs are indicative of phosphoproteins, a search was performed on various known phosphoproteins using quantitative pattern matching algorithms (Appendix A). These phosphoproteins were arbitrarily chosen mostly from the PhosphoBase database. To decrease the amount of sequence variation, the selected proteins are all human. For pattern matching, the original Blocks database (http://blocks.fhcrc.org/blocks/blocks_search.html) and the eMatrix database (http://ematrix.stanford.edu) were used to identify known motifs which may be indicative of phosphorylation. The Blocks database was used because it offered the most detailed results, much of which may be false or statistically insignificant, while the eMatrix database, in contrast, provides fewer statistical details but more refined results (Wu et al, 1999).
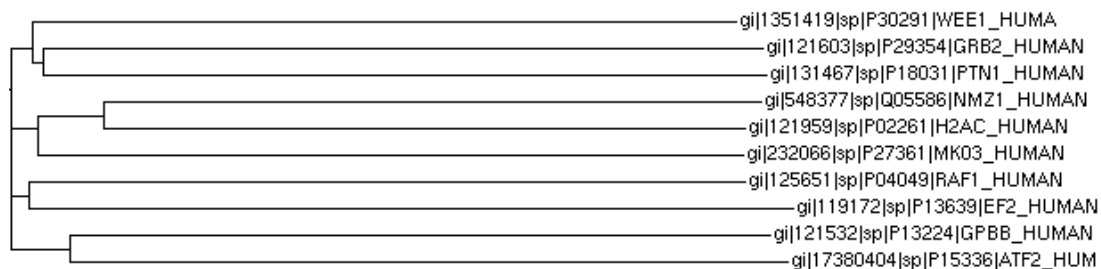
The purpose of using quantitative pattern matching was to possibly identify known motifs common to phosphorylated proteins. Perhaps documented motifs existed, initially attributed to other protein features, which may be indicative of phosphorylation as well. (For instance, perhaps these databases reveal that all phosphoproteins have a helix-loop-helix motif.) However, neither of the original BLOCKS database nor eMatrix identified any known motifs which are common to most or all of the phosphoproteins. While the eMatrix program was successful at identifying various biologically relevant features of the phosphoproteins (for example, identifying the "Raf-like Ras-binding domain" from the human Raf sequence), it found no unifying motifs between all phosphoproteins when limited by either a 0.01 or a 0.1 significance level. The eMatrix program did find some similarities between certain small pairs/subsets of phosphoproteins. For instance, the MAP kinase signature was found in the Wee1, p44/Erk1, and Raf sequences, confirming the acknowledged generalization that MAP kinases are phosphorylated.

Although the pattern matching databases were unable to discover any motif common to all ten phosphoproteins searched, that does not necessarily mean that no such motif(s) exist, but merely that such proposed motifs have yet to be identified and entered into the BLOCKS database. The next logical step in identifying potential motifs common to phosphoproteins was to perform sequence alignments. The ten sequences were entered into the ClustalW program available on the Decypher machine. This multiple sequence alignment did not produce any regions of homology. This result was not entirely unexpected given that these phosphoproteins are so different from each other in many respects. These differences would likely provide a noisy negative background signal that

would drown out a potential "phosphoprotein" motif, which may only be several residues long. Lowering the gap opening / extension penalties from 10 / 0.05 to 1 / 0.005 did not help significantly, as the latter criteria only provided single amino acid alignments rather than the desired motifs of multiple contiguous residues.

In an attempt to maximize the likelihood that motifs are uncovered from a multiple sequence alignment, pairs of phosphoproteins most likely to have evolutionary similarity, as determined by the dendrogram from the previous ClustalW alignment (Figure 4), were aligned via ClustalW. As expected, more residues were aligned when dealing with two sequences rather than ten sequences. However, no clear motifs emerged from these alignments either. For example, when the phosphoproteins Grb2 and Ptn1 were aligned with ClustalW with gap penalties of 10 / 0.05, thirty-nine amino acids were aligned identically between the two sequences. But of these 39 residues, there existed only two pairs of consecutive identical amino acids, and no three consecutive amino acids were identical in both sequences. There exist several regions where there are short stretches of similar, though not identical sequences (i.e. stretches with conservative substitutions such as phenylalanine to tryptophan, both large hydrophobic groups); motifs common to phosphoproteins may be hidden within these alignments.

Figure 4: Clustal W dendogram of phosphoproteins

Conclusions

With the prevalence of phosphorylation as a mechanism for controlling various cellular functions, there exists a need to identify whether certain proteins are subject to this posttranslational modification. Identifying novel phosphoproteins and phosphorylation sites would help clarify the function of previously unclassified proteins. Insights about a protein's phosphorylation state shed light upon that protein's place in a biological pathway.

Currently, only one bioinformatics resource is available to identify potential phosphorylation sites from amino acid sequences. This program, NetPhos, relies on an artificial neural networked trained to recognize phosphorylated sequences. While this program does narrow down the number of serines/threonines/tyrosines likely to be phosphorylated, it tends to provide an overly tolerant threshold, such that many false positives result from a NetPhos sequence search. Furthermore, NetPhos only considers residues in close proximity to the phosphate acceptor, ignoring possible motifs lying further away from the acceptor residue as well as secondary structure motifs which may predict phosphorylation.

Using a combination of pattern matching and sequence alignment, this research attempted to find novel methods of predicting phosphorylation by identifying possible motifs indicative of phosphorylation. No discernable pattern was found using the numerous variants of the BLOCKS database, meaning that any potential phosphorylation-motif had yet to be classified. Several multiple sequence alignments were subsequently performed on phosphoprotein motifs, attempting to identify any sequence homology between phosphoproteins which could be candidate "phospho-

motifs." Overall, the sequence alignments proved unable to produce any obvious motifs, though a more exhaustive search may yet uncover significant regions of similarity between phosphoproteins.

While motif(s) which signify phosphorylation have yet to be unearthed, the fact remains that newer algorithms of predicting phosphorylation are necessary to account for the three-dimensional nature of phosphorylation events between kinases and substrate phosphoproteins. Combinations of existing bioinformatics tools, as presented here, or novel methodologies may provide better predictors of sequences inclined to be phosphorylated.

References:

Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999 Dec 17;294(5):1351-62.

Kreegipuu A, Blom N, Brunak S. PhosphoBase, a database of phosphorylation sites: release 2.0. Nucleic Acids Res. 1999 Jan 1;27(1):237-9.

Nakai K. Review: prediction of in vivo fates of proteins in the era of genomics and proteomics. J Struct Biol. 2001 May-Jun;134(2-3):103-16.

Pawson T. Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. Cell. 2004 Jan 23;116(2):191-203.

Wu TD, Nevill-Manning CG, Brutlag DL. Minimal-risk scoring matrices for sequence analysis. J Comput Biol. 1999 Summer;6(2):219-35.

Appendix 1: Phosphoprotein sequences in FASTA format

>gi|1351419|sp|P30291|WEE1_HUMAN Wee1-like protein kinase (WEE1hu)
MSFLSRQQPPPPRRAGAACTLRQKLIFSPCSDCEEEEEEEEEEEGSGHSTGEDSAFQEPDSPLPPARSPTE
PGPERRRSPGPAPGSPGELEEDLLLPGACPGADEAGGGAEGDSWEEEGFGSSSPVKSPAAPYFLGSSFSP
VRCGGPGDASPRGCGARRAGEGRRSPRPDHPGTPPHKTFRKLRLFDTPHTPKSLLSKARGIDSSSVKLRG
SSLFMDTEKSGKREFDVRQTPQVNINPFTPDSLLLHSSGQCRRRKRTYWNDSCGEDMEASDYELEDETRP
AKRITITESNMKSRYTTEFHELEKIGSGEFGSVFKCVKRLDGCYAIKRSKKPLAGSVDEQNALREVYAH
AVLGQHSHVVRYFSAWAEDDHMLIQNEYCNGGSLADAISENYRIMSYFKEAELKDLLLQVGRGLRYIHSM
SLVHMDIKPSNIFISRTSIPNAASEEGDEDDWASNKVMFKIGDLGHVTRISSPQVEEGDSRFLANEVLQE
NYTHLPKADIFALALTVVCAAGAEPLPRNGDQWHEIRQGRLPRIPQVLSQEFTELLKVMIHPDPERRPSA
MALVKHSVLLSASRKSAEQLRIELNAEKFKNSLLQKELKKAQMAKAAAEERALFTDRMATRSTTQSNRTS
RLIGKKMNRSVSLTIY

>gi|121603|sp|P29354|GRB2_HUMAN Growth factor receptor-bound protein 2 (GRB2 adapter protein) (SH2/SH3 adapter GRB2) (ASH protein)
MEAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIPKNYIEMKPHPWFFGKIPRAKA
EEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYFLWVVKFNSLNELVDYHRSTSV
SRNQQIFLRDIEQVPQQPTYVQALFDFDPQEDGELGFRRGDFIHVMDNSDPNWWKGACHGQTGMFPRNYV
TPVNRNV

>gi|125651|sp|P04049|RAF1_HUMAN RAF proto-oncogene serine/threonine-protein kinase (Raf-1) (C-RAF) (cRaf)
MEHIQGAWKTISNGFGFKDAVFDGSSCISPTIVQQFGYQRRASDDGKLTDPSKTSNTIRVFLPNKQRTVV
NVRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWNTDAASLIGEELQVDFLDHVPLTTHN
FARKTFLKLAFCDICQKFLLNGFRCQTCGYKFHEHCSTKVPTMCVDWSNIRQLLLFPNSTIGDSGVPALP
SLTMRRMRESVSRMPVSSQHRYSTPHAFTFNTSSPSSEGSLSQRQRSTSTPNVHMVSTTLPVDSRMIEDA
IRSHSESASPSALSSSPNNLSPTGWSQPKTPVPAQRERAPVSGTQEKNKIRPRGQRDSSYYWEIEASEVM
LSTRIGSGSFGTVYKGKWHGDVAVKILKVVDPTPEQFQAFRNEVAVLRKTRHVNILLFMGYMTKDNLAIV
TQWCEGSSLYKHLHVQETKFQMFQLIDIARQTAQGMDYLHAKNIIHRDMKSNNIFLHEGLTVKIGDFGLA
TVKSRWSGSQQVEQPTGSVLWMAPEVIRMQDNNPFSFQSDVYSYGIVLYELMTGELPYSHINNRDQIIFM
VGRGYASPDLSKLYKNCPKAMKRLVADCVKKVKEERPLFPQILSSIELLQHSLPKINRSASEPSLHRAAH
TEDINACTLTTSPRLPVF

>gi|548377|sp|Q05586|NMZ1_HUMAN Glutamate [NMDA] receptor subunit zeta 1 precursor (NR1)
MSTMRLLTLALLFSCSVARAACDPKIVNIGAVLSTRKHEQMFREAVNQANKRHGSWKIQLNATSVTHKPN
AIQMALSVCEDLISSQVYAILVSHPPTPNDHFTPTPVSYTAGFYRIPVLGLTTRMSIYSDKSIHLSFLRT
VPPYSHQSSVWFEMMRVYSWNHIILLVSDDHEGRAAQKRLETLLEERESKAEKVLQFDPGTKNVTALLME
AKELEARVIILSASEDDAATVYRAAAMLNMTGSGYVWLVGEREISGNALRYAPDGILGLQLINGKNESAH
ISDAVGVVAQAVHELLEKENITDPPRGCVGNTNIWKTGPLFKRVLMSSKYADGVTGRVEFNEDGDRKFAN
YSIMNLQNRKLVQVGIYNGTHVIPNDRKIIWPGGETEKPRGYQMSTRLKIVTIHQEPFVYVKPTLSDGTC
KEEFTVNGDPVKKVICTGPNDTSPGSPRHTVPQCCYGFCIDLLIKLARTMNFTYEVHLVADGKFGTQERV
NNSNKKEWNGMMGELLSGQADMIVAPLTINNERAQYIEFSKPFKYQGLTILVKKEIPRSTLDSFMQPFQS
TLWLLVGLSVHVVAVMLYLLDRFSPFGRFKVNSEEEEEDALTLSSAMWFSWGVLLNSGIGEGAPRSFSAR
ILGMVWAGFAMIIVASYTANLAAFLVLDRPEERITGINDPRLRNPSDKFIYATVKQSSVDIYFRRQVELS
TMYRHMEKHNYESAAEAIQAVRDNKLHAFIWDSAVLEFEASQKCDLVTTGELFFRSGFGIGMRKDSPWKQ
NVSLSILKSHENGFMEDLDKTWVRYQECDSRSNAPATLTFENMAGVFMLVAGGIVAGIFLIFIEIAYKRH
KDARRKQMQLAFAAVNVWRKNLQDRKSGRAEPDPKKKATFRAITSTLASSFKRRRSSKDTSTGGGRGALQ
NQKDTVLPRRAIEREEGQLQLCSRHRES

>gi|121959|sp|P02261|H2AC_HUMAN Histone H2A.c/d/i/n/p (H2A.1) (H2A/c) (H2A/d) (H2A/i) (H2A/n) (H2A/p) (H2A.1b)
MSGRGKQGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYAERVGAGAPVYLAAVLEYLTAEILELAGNA
ARDNKKTRIIPRHLQLAIRNDEELNKLLGKVTIAQGGVLPNIQAVLLPKKTESHHKAKGK

>gi|121532|sp|P13224|GPBB_HUMAN Platelet glycoprotein Ib beta chain precursor (GP-Ib beta) (GPIbB) (GPIb-beta) (CD42B-beta) (CD42C)
MGSGPRGALSLLLLLLAPPSRPAAGCPAPCSCAGTLVDCGRRGLTWASLPTAFPVDTTELVLTGNNLTAL
PPGLLDALPALRTAHLGANPWRCDCRLVPLRAWLAGRPERAPYRDLRCVAPPALRGRLLPYLAEDELRAA
CAPGPLCWGALAAQLALLGLGLLHALLLVLLLCRLRRLRARARARAAARLSLTDPLVAERAGTDES

>gi|17380404|sp|P15336|ATF2_HUMAN Cyclic-AMP-dependent transcription factor ATF-2 (Activating transcription factor 2) (cAMP response element binding protein CRE-BP1) (HB16)
MSDDKPFLCTAPGCGQRFTNEDHLAVHKHKHEMTLKFGPARNDSVIVADQTPTPTRFLKNCEEVGLFNEL
ASPFENEFKKASEDDIKKMPLDLSPLATPIIRSKIEEPSVVETTHQDSPLPHPESTTSDEKEVPLAQTAQ
PTSAIVRPASLQVPNVLLTSSDSSVIIQQAVPSPTSSTVITQAPSSNRPIVPVPGPFPLLLHLPNGQTMP
VAIPASITSSNVHVPAAVPLVRPVTMVPSVPGIPGPSSPQPVQSEAKMRLKAALTQQHPPVTNGDTVKGH
GSGLVRTQSEESRPQSLQQPATSTTETPASPAHTTPQTQSTSGRRRRAANEDPDEKRRKFLERNRAAASR
CRQKRKVWVQSLEKKAEDLSSLNGQLQSEVTLLRNEVAQLKQLLLAHKDCPVTAMQKKSGYHTADKDDSS
EDISVPSSPHTEAIQHSSVSTSNGVSSTSKAEAVATSVLTQMADQSTEPALSQIVMAPSSQSQPSGS

>gi|131467|sp|P18031|PTN1_HUMAN Protein-tyrosine phosphatase, non-receptor type 1 (Protein-tyrosine phosphatase 1B) (PTP-1B)
MEMEKEFEQIDKSGSWAAIYQDIRHEASDFPCRVAKLPKNKNRNRYRDVSPFDHSRIKLHQEDNDYINAS
LIKMEEAQRSYILTQGPLPNTCGHFWEMVWEQKSRGVVMLNRVMEKGSLKCAQYWPQKEEKEMIFEDTNL
KLTLISEDIKSYYTVRQLELENLTTQETREILHFHYTTWPDFGVPESPASFLNFLFKVRESGSLSPEHGP
VVVHCSAGIGRSGTFCLADTCLLLMDKRKDPSSVDIKKVLLEMRKFRMGLIQTADQLRFSYLAVIEGAKF
IMGDSSVQDQWKELSHEDLEPPPEHIPPPPRPPKRILEPHNGKCREFFPNHQWVKEETQEDKDCPIKEEK
GSPLNAAPYGIESMSQDTEVRSRVVGGSLRGAQAASPAKGEPSLPEKDEDHALSYWKPFLVNMCVATVLT
AGAYLCYRFLFNSNT

>gi|119172|sp|P13639|EF2_HUMAN Elongation factor 2 (EF-2)
MVNFTVDQIRAIMDKKANIRNMSVIAHVDHGKSTLTDSLVCKAGIIASARAGETRFTDTRKDEQERCITI
KSTAISLFYELSENDLNFIKQSKDGAGFLINLIDSPGHVDFSSEVTAALRVTDGALVVVDCVSGVCVQTE
TVLRQAIAERIKPVLMMNKMDRALLELQLEPEELYQTFQRIVENVNVIISTYGEGESGPMGNIMIDPVLG
TVGFGSGLHGWAFTLKQFAEMYVAKFAAKGEGQLGPAERAKKVEDMMKKLWGDRYFDPANGKFSKSATSP
EGKKLPRTFCQLILDPIFKVFDAIMNFKKEETAKLIEKLDIKLDSEDKDKEGKPLLKAVMRRWLPAGDAL
LQMITIHLPSPVTAQKYRCELLYEGPPDDEAAMGIKSCDPKGPLMMYISKMVPTSDKGRFYAFGRVFSGL
VSTGLKVRIMGPNYTPGKKEDLYLKPIQRTILMMGRYVEPIEDVPCGNIVGLVGVDQFLVKTGTITTFEH
AHNMRVMKFSVSPVVRVAVEAKNPADLPKLVEGLKRLAKSDPMVQCIIEESGEHIIAGAGELHLEICLKD
LEEDHACIPIKKSDPVVSYRETVSEESNVLCLSKSPNKHNRLYMKARPFPDGLAEDIDKGEVSARQELKQ
RARYLAEKYEWDVAEARKIWCFGPDGTGPNILTDITKGVQYLNEIKDSVVAGFQWATKEGALCEENMRGV
RFDVHDVTLHADAIHRGGGQIIPTARRCLYASVLTAQPRLMEPIYLVEIQCPEQVVGGIYGVLNRKRGHV
FEESQVAGTPMFVVKAYLPVNESFGFTADLRSNTGGQAFPQCVFDHWQILPGDPFDNSSRPSQVVAETRK
RKGLKEGIPALDNFLDKL

>gi|232066|sp|P27361|MK03_HUMAN Mitogen-activated protein kinase 3 (Extracellular signal-regulated kinase 1) (ERK-1) (Insulin-
stimulated MAP2 kinase) (MAP kinase 1) (MAPK 1) (p44-ERK1) (ERT2) (p44-MAPK) (Microtubule-associated protein-2 kinase)
MAAAAAQGGGGGGEPRRTEGVGPGVPGEVEMVKGQPFDVGPRYTQLQYIGEGAYGMVSSAYDHVRKTRVAI
KKISPFEHQTYCQRTLREIQILLRFRHENVIGIRDILRASTLEAMRDVYIVQDLMETDLYKLLKSQQLSN
DHICYFLYQILRGLKYIHSANVLHRDLKPSNLLINTTCDLKICDFGLARIADPEHDHTGFLTEYVATRWY
RAPEIMLNSKGYTKSIDIWSVGCILAEMLSNRPIFPGKHYLDQLNHILGILGSPSQEDLNCIINMKARNY
LQSLPSKTKVAWAKLFPKSDSKALDLLDRMLTFNPNKRITVEEALAHPYLEQYYDPTDEPVAEEPFTFAM
ELDDLPKERLKELIFQETARFQPGVLEAP