Published in final edited form as:

Proteins. 2009 February 15; 74(3): 655–668. doi:10.1002/prot.22181.

Cell cycle kinases predicted from conserved biophysical properties

Kazimierz O. Wrzeszczynski^{1,2,3,4} and Burkhard Rost^{1,2,3,*}

¹ Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, USA

² Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA

³ NorthEast Structural Genomics Consortium (NESG), Columbia University, 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA

⁴ Integrated Program in Cellular, Molecular, Structural and Genetic Studies, Columbia University, 630 West 168th Street, New York, NY 10032, USA

Abstract

Machine learning techniques can classify functionally related proteins where homology-transfer as well as sequence and structure motifs fail. Here, we present a method that aimed at complementing homology-transfer in the identification of cell cycle control kinases from sequence alone. First, we identified functionally significant residues in cell cycle proteins through their high sequence conservation and biophysical properties. We then incorporated these residues and their features into support vector machines (SVM) to identify new kinases and more specifically to differentiate cell cycle kinases from other kinases and other proteins. As expected, the most informative residues tend to be highly conserved and tend to localize in the ATP binding regions of the kinases. Another observation confirmed that ATP binding regions are typically not found on the surface but in partially buried sites, and that this fact is correctly captured by accessibility predictions. By using these highly conserved, semi-buried residues and their biophysical properties we could distinguish cell cycle S/T kinases from other kinase families at levels around 70–80% accuracy and 61%–81% coverage. An application to the entire human proteome predicted at least 97 human proteins with limited previous annotations to be candidates for cell cycle kinases.

Keywords

cell cycle kinases; serine/threonine kinase; ATP binding; protein function prediction; homology transfer; sequence conservation; functional residues; support vector machine

Introduction

Our major overall objective is the development of methods that identify all human proteins involved in cell cycle control by combining all available data with *in silico* predictions^{1,2}. One task to solve toward that aim is the development of an automatic method that specifically distinguishes between those kinases that are involved in cell cycle control and those that are

*Corresponding author: kow2001@columbia.edu, <http://www.rostlab.org/>, Tel: +1-212-851-4669.

not. This goal bears two questions that are intrinsically difficult to answer. (1) What exactly is experimentally known? (2) What exactly can be done through homology-based inferences?

Experimental evidence is increasingly statistical in nature

The complete annotation of protein structure and function for all proteins in e.g. human is a highly important challenge for modern medical and molecular biology. Despite rapid advances in experimental high-throughput techniques, the gap between sequence and annotation has been exploding ever since the onset of the genome era³. The most successful bridge for this sequence/function gap comes from computational biology applying homology-transfer, i.e. the inference of experimental information experimentally collected for model organisms^{3–7}. With the advent of high-throughput experiments, the line between *solid experimental evidence* and *statistical computational prediction* increasingly gets blurred. For our objective to automatically distinguish cell cycle kinases from other kinases, the problem becomes which experimental information can we consider as a solid standard-of-truth. We approximated a solution by curating a data set of cell cycle proteins through database lookups and reading the literature¹. Obviously, such an approach has to remain preliminary, but it was the best we could do at the moment. Our curated set provides biophysical protein features in leading toward cell cycle kinase classification.

Homology-transfer and de novo predictions overlap

Another line that is blurred is the one between homology-transfer (infer experimental data from a mouse protein to its human homologue) and *de novo* prediction (predict function for a human protein without any experimental information for recognizable homologues). Firstly, the levels of sequence similarity that make automatic homology-transfer succeed depend crucially on the precise feature of function to be transferred^{8–13}. For instance, the observation that two proteins interact in mouse by no means implies the homologous pair interacts in human¹². Similarly, two homologous proteins may respond very differently to single nucleotide polymorphisms¹⁴; they may slightly move their interaction hotspots^{15–17}, and they may definitely differ substantially in the way they control the cell cycle¹⁸. Secondly, depending on the sequence similarity, *de novo* predictions may be more accurate than homology-transfer^{19–22}.

___DELETION___ Our ultimate goal of specifically identifying cell cycle kinases can extend beyond the level of detail of earlier homology-based work¹. For instance, protein phosphorylation pathways are a vital component in almost every cellular processes²³ but the EC classification²⁴ of all protein kinases has been expanded from one single “protein kinase” class (2.7.1.37) only in 2005 to specifically include individual kinases and their substrates (2.7.11.x; where x originally included 12 subclasses and now recently has expanded further to 31 subclasses).

Methods that characterize cell cycle kinases

Computational methods examining the cell cycle focus heavily on genetic and protein data to build cellular networks²⁵. Surprisingly few methods explicitly analyze the biochemical nature of cell cycle proteins despite the large amount of available data^{2,23,26–28}. Brunak and co-workers have been most active in this field expanding on their efforts from earlier machine-learning based identification of cell cycle proteins of the yeast proteome to more currently the analysis of conservation for transcriptional and post-translation regulation of cell cycle proteins in several proteomes^{18,29}.

Homology-transfer of function annotation is even more challenging¹⁰. Advances in function annotation have relied on highly conserved catalytic residues^{23,30–34}. Identifying and correctly aligning catalytic residues can lead toward a prediction of function precise enough to distinguish between closely related proteins^{23,35,36} and has been the accepted operational solution to classify kinases^{37,38}. Functionally related proteins still often vary greatly with

regard to substrate and cellular pathway within kinases^{23,39,40}; therefore structurally similar and homologous proteins may have very different roles within the cell.

Elegant and detailed studies of enzymatic mechanisms demonstrate how minor but substantial changes in the active site can alter substrate specificity and function^{26,41–46}. The ability to capture these differences of kinase enzymatic mechanisms should allow for a more robust classification of kinases involved in various cellular pathways. We have previously analyzed proteins involved in cell cycle control and the ability to transfer cell cycle control function through homology¹. We used this detailed level of homology-transfer to build accurate profiles of cell cycle kinases identifying highly conserved functional residues. Biophysical information from these residues was used as the main feature towards the classification of cell cycle function among kinases.

Here, we described a novel machine learning-based method that predicts protein function at a very detailed level, namely the classification specific to a particular cellular process, the cell cycle. To some degree, the method identified the particular enzymatic specificity among kinases through implicitly finding the importantly conserved residues. Our results can benefit annotations for both a biological process and a cellular process.

Methods

Cell cycle control data

We searched for proteins with trusted experimental information that established their involvement in the cell cycle process in SWISS-PROT^{47,48}. Most of the proteins retrieved in this manner control the G1/S and G2/M transitions, or are related to the M and S phases. In total, we found 595 proteins; 97 of these had multiple annotations and 229 had some “other” annotation than the four phase classification involving the cell cycle process.

DELETION In order to reduce the bias from too similar sequences, we generated sequence-unique subsets for all types of proteins under consideration. *Sequence-unique* was defined by having no pair in the set with an HSP-value>0 (HVAL, Eqn. 1 below) which roughly corresponded to <33% identical residues over more than 100 aligned residues. Given an all-against-all pairwise alignment for the biased set, we simply used a greedy search to find the largest subset that fulfilled the above condition⁴⁹. This reduced the entire set to 119 sequence-unique cell cycle proteins.

Non-cell cycle proteins

As a negative set for the homology-transfer analysis we expanded on our dataset from our previous study¹ and used 15,192 proteins with SWISS-PROT annotations that indicated functions other than cell cycle (these were sequence-unique at HVAL=5, Eqn. 1). In particular, we included proteins from all EC classes, all subcellular localizations and proteins annotated as non-enzymes.

Kinase data

Using the SWISS-PROT and kinase.com databases, we identified a total of 10,718 multi-species protein kinase sequences. 534 of these were annotated as cell cycle kinases, and 268 of the 534 were Serine/Threonine kinases (S/T kinases). At HVAL=15 (Eqn. 1) we retained 29 sequence-unique cell cycle S/T kinases. These 29 S/T kinases proteins constituted our main set of positives (part of cell cycle) for all classifications. The negatives contained 1755 (134 sequence-unique at HVAL<15) non-cell cycle kinases that included 480 tyrosine kinases (33 sequence-unique at HVAL<15). Human protein sequences used for classification were obtained from NCBI Reference Sequence (RefSeq) Database Release 22 (ftp://ftp.ncbi.nih.gov/refseq/) ⁵⁰.

Sequence alignment

We applied pairwise BLAST⁵¹ to align all 595 proteins from our trusted cell cycle dataset against a unique set of proteins known not to be related with cell cycle (total of 15,192 proteins which included proteins from all EC classes, all subcellular localizations and proteins annotated as non-enzymes). We refined the analysis by generating PSI-BLAST⁵¹ profiles using a standard procedure with five iterations⁵². These profiles were then aligned against all proteins used in our study. ___DELETION___

Sequence similarity

We used three different measures for sequence similarity. (1) PIDE: The simplest was the percentage pairwise sequence identity (PIDE), i.e. the percentage of identical residues ignoring insertions and deletions (e.g. KK=1, KR=0). (2) HVAL: Aligning two stretches of 5 residues at 100% PIDE may be totally meaningless, while 50% PIDE over 200 residues may be very meaningful, i.e. the length of the alignment is an important criterion when measuring sequence similarity. The HSSP-value (HVAL) captures sequence similarity by relating the percentage of identical residues (PID) to the length of an alignment (L). The underlying empirical threshold is the HSSP-curve optimized for the detection of structural similarity^{8,49,53}. Positive HSSP-values (HVAL>0) imply structural similarity (the larger the number, the more similar the two proteins), while negative HSSP-values (HVAL<0) imply that the structural similarity cannot be inferred from pairwise sequence comparisons, i.e. estimate the distance between two proteins (the more negative, the more distant the two proteins). For instance, HVAL=0 implies 33% PID for 100 aligned residues, and about 21% for 250 aligned residues. More precisely, the HSSP-value (HVAL) is defined by:

$$\text{HVAL} = \text{PID} - \begin{cases} 100 & \text{for } L \leq 11 \\ 480 \cdot L^{-0.32 \cdot \{1 + \exp(-L/1000)\}} & \text{for } L \leq 450 \\ 19.5 & \text{for } L > 450 \end{cases} \quad (\text{Eqn. 1})$$

(3) EVAL: p- and E-values measure the probability for achieving a certain level of sequence similarity given some assumptions about the background. In particular in the context of using sequence-profile alignments derived from PSI-BLAST⁵¹, the use of E-values becomes an essential measure for sequence similarity. We used the values directly from PSI-BLAST. The HSSP-value is optimized to capture structural similarity; nevertheless, it also accurately reflects similarity in terms of various aspects of protein function^{1,13}. We used the HSSP-value to eyeball a threshold in similarity that was used to separate data sets, namely HVAL=10, this implied that the accuracy of homology-transfer was below 80% and that only 40% of the true positives (cell cycle proteins) could be identified by homology-transfer (Fig. 2). In order to be even more conservative, we separated our data sets such that no pair of proteins used for training and testing had a similarity of HVAL>0 (at that point homology-transfer was less than 60% accurate, Fig. 2).

Measuring performance

We used the following definition to measure accuracy (often referred to as specificity):

$$\text{Accuracy} = 100 \cdot \frac{\text{number of true pairs found above threshold}}{\text{number of all pairs above threshold}} \quad (\text{Eqn. 2})$$

with the thresholds given by either of three measures for sequence similarity (PIDE, HVAL, EVAL). We considered all pairs as *true* that were experimentally found in our trusted cell cycle

data set¹. In analogy, we used the following definitions for coverage (often referred to as sensitivity):

$$\text{Coverage} = 100 \cdot \frac{\text{number of true pairs found above threshold}}{\text{number of all true pairs}} \quad (\text{Eqn. 3})$$

Highly conserved residues: 65% level

We generated PSI-BLAST profiles for each protein in our data set by searching against a *non-redundant* (filtered at 95%) subset of SWIS-PROT + TrEMBL + PDB + our cell cycle data set. A final PSI-BLAST E-value of 10^{-10} (-e, PSI-BLAST parameter) was selected in all profiles for conserved residue selection. One other PSI-BLAST parameter (-b 500) was optimized to achieve high levels of accuracy. We then defined as “conserved” any residue that was observed in more than $\geq 65\%$ of all proteins aligned at $\text{EVAL} < 10^{-10}$.

Structure prediction

We used PROFsec^{54–57} and PROFacc^{54,56–58} to predict secondary structure and solvent accessibility, respectively. As input for our SVMs we retained only the “binary” states predicted by those methods, i.e. H (helix), E (strand), L (other), b (buried), i (intermediate) and e (exposed).

Solvent accessibility ATP and GTP binding sites

We collected all PDB⁵⁹ entries (available as of April 2007) that contained the terms *ATP* or *GTP* in the HETNAM field. Using a similar approach we previously used for identifying DNA contacting residues⁶⁰, amino acids and ligands were considered in contact if any of their atoms were closer than 4 Å. As for our cell cycle data, we predicted secondary structure by PROFsec^{54–57} and accessibility by PROFacc^{54,56–58}. Experimental values for accessibility were taken from DSSP⁶¹.

SVM training

SVM is a machine learning method based on statistical learning theory⁶² that has been used extensively for problems in computational biology as diverse as text analysis⁶³, identification of distant structural relations⁶⁴, prediction of subcellular localization^{22,65}, the sequence based prediction of druggable proteins⁶⁶ and detection of protein-coding regions in RNA sequences⁶⁷. We used the Gist 2.3 SVM package⁶⁸. **__DELETION__** In training, we maintained a ratio between positives (cell cycle) and negatives (non-cell cycle) of 1:2 (positive:negative). We trained with the standard radial basis function kernel, and optimized other free parameters once and for all optimizing accuracy values on a limited subset of the training data. Training data was obtained from only one two-fold split in the cross-validation process (e.g. for 29 positive samples, 28 are used for training and one is left out and training did not revolve for all remaining split possibilities while testing did). Once the final training set parameters were established all cross-validation (the entire passes through the leave one out protocol) was performed under one set of final and consistent parameters for the full positive and varying (50X) negative data sets. A two-norm soft margin, a Gist radial basis kernel width factor between 2 to 3 and adjusting the Gist constant near the default 10 while accompanying all other Gist default parameters was selected for best performance.

Input

For each protein we use the following information as input to the SVM: amino acid composition (20 units), amino acid solvent exposure in three states (60 units), amino acid secondary structure in three states (60 units). These units were obtained from the entire protein sequence.

The standard challenge for building evolutionary profiles that describe families of related proteins is to find a good compromise between specificity (often referred to as *accuracy*) and selectivity (also referred to as *coverage*). We used our estimates for the performance of homology-transfer (Fig. 2) to constrain the construction of profiles. In particular, we included only proteins for which homology-transfer remained around 75–80% accuracy (Fig. 2).

We then selected the same features from the subset of the conserved residues. We considered residues as conserved when one particular amino acid was conserved at that position for over 65% of all the family members. This threshold was chosen to optimize performance on a subset of the training set (Fig. 1). Therefore if an amino acid is not highly conserved at accurate levels of homology transfer (Fig. 1) we devalued its contribution to the overall features used as input. With this protocol we hoped to focus on the residues with the most significant functional and structural information.

Cross-validation

Nowhere in this manuscript do we refer to the performance on the training set. Instead, all our estimates for performance were valid for test sets that were generated by the following rigorous protocol for cross-validation. First, we clustered our data into groups that were such that homology-transfer might be possible within but not between groups. Second, we performed a leave-one-out test of the type that used one protein from one group for testing, and the proteins from all other groups for training. Note that due to the fact that we had so many more negatives than positives, we repeated this process 50 times for each test protein; each time, we chose a different random subset of proteins as the negatives for training. In particular, we chose 50 different negative data sets (containing random negative examples from our much larger sequence unique negative data set). The error estimates for accuracy and coverage were based on the 50 multiple runs of the established training protocol with varying negative test sets maintaining positives the same. The standard deviation (σ) for both accuracy and coverage was computed from the average performance after testing with each new negative data set.

Human Kinase Prediction

Each genome sequence from the human genome NCBI Reference Sequence (RefSeq) Database Release 22 underwent the same protocol of our testing and training sets. Sequence profiles and PROFac predictions of each human protein sequence were established and conserved residues and their biophysical features were identified. The conserved residues and their biophysical properties were used in classification. Each protein was tested against our multiple ($n=50$) trained sample weights and an average classification discriminant from the Gist output was calculated. A sequence was predicted as positive (implying that it was predicted to be part of the cell cycle) if the discriminant average after 50 SVM iterations was >0.25 (with 1.00 being set as a true positive cell cycle function discriminant value in our training and testing sets). A discriminant of ≥ 0.25 provided 685 predictions while a more conservative discriminant value of ≥ 0.50 resulted in 170 predictions (the 170 sequences are provided in the supplementary material).

Results

Thresholds for reliable homology transfer

If we want to find proteins that have similar roles in the cell cycle as the proteins for which we have experimental information in public databases, we have to first establish a threshold for *significant sequence similarity*, i.e. we have to address the question: at which level of sequence similarity can we infer similarity in the specific functional role of that protein. Such thresholds have to find a balance between accuracy (number of true pairs – protein pairs with similar function - found above threshold/number of all pairs above threshold) and coverage (number

of true pairs above threshold/number of all true pairs). We found that at $HVAL > 40$ (Eqn. 1; ~73% pairwise sequence identity at 100 aligned residues), we could seemingly infer the role in the cell cycle at an accuracy of 93% (Fig. 2); decreasing to $HVAL > 10$ yielded 78% accuracy. Comparable thresholds in PSI-BLAST E-values were: 10^{-10} (80% accuracy) and 10^{-25} (89% accuracy). Cutting at PSI-BLAST E-values of $< 10^{-3}$ yielded 29% accuracy. The success of the HVAL (Eqn. 1) was demonstrated by e.g. comparing the thresholds in similarity needed to reach about 80% accuracy, these were reached by $HVAL > 10$ and by percentage pairwise sequence identity (PIDE) of 64% (Fig. 2). As $HVAL > 10$ corresponds to something like $PIDE > 43\%$ instead of 64%, HVAL cutoffs gave much higher coverage at the same accuracy. Note in particular that the EVAL of 10^{-3} routinely applied when filtering PSI-BLAST alignments distinguished cell cycle proteins very poorly.

Evolutionary information used rather restrictively

We effectively devalued the contribution from non-conserved residues (Fig. 1, Methods). Two parameters predominantly steered the diversity of the evolutionary information used as input, namely the PSI-BLAST Eval (choice: $< 10^{-10}$), and threshold for the minimal fraction of family members required to have the same amino acid at a given position (choice: > 0.65). With these two choices, we retained on average about $1/4^{\text{th}}$ of the protein sequence (Fig. 3) as effective input (as not all residues are buried and conserved). A more restrictive threshold at $EVAL < 10^{-100}$ resulted in a similar fraction of the protein's residues included (Fig. 3). Overall, our approach effectively favored the compilation of profiles only for very similar proteins.

Functional residues predicted as buried

When isolating highly conserved functional residues from our positive (cell cycle) and negative (non-cell cycle) data sets, we noticed that over 67% of the conserved residues were predicted as either having buried or intermediate solvent accessibility. Most ATP binding residues in cyclin-dependent kinases from our cell cycle data set were also predicted not to be exposed. We therefore examined all PROFacc⁵⁶ predictions for ATP and GTP binding sites in the PDB, and found that over 80% of the ATP and GTP binding sites were predicted as buried or intermediately accessible⁵⁸ (Fig. 4A). Assigning accessibility directly from the experimental structures by DSSP⁶¹ confirmed this as a generic result as opposed to as a prediction mistake (Fig. 4B). This result confirmed what experts expected and what particular detailed studies showed^{69–72}.

Classification of protein kinases involved in the cell cycle process

If we accept 13% error rate and want to differentiate at least 41% of all known cell cycle S/T kinases from all known tyrosine kinases, we need to threshold our homology-transfer at levels $> 60\%$ pairwise sequence identity (Table 1). At this level we reach roughly the same accuracy/coverage (85%/41%, Table 1) in distinguishing S/T cell cycle from S/T non-cell cycle kinases. Using SVMs to aid homology-transfer (Methods), we attempted to specifically identify cell cycle kinases from among three different sets of *negatives* or baits, namely (i) from non-cell cycle tyrosine kinases (30 unique in a set of 268), (ii) from non-cell cycle serine/threonine (S/T) kinases (134 unique of 1808), and (iii) from a mix of proteins with function annotations other than cell cycle kinase (2058 unique). We tested SVMs with different types of input information (Methods). Each resulting solution (SVM) outperformed the corresponding homology-transfer (Fig. 5).

Human kinome as control

We wanted to test our protocol against a current standard of known kinases. As a control experiment we trained SVMs with the same types of inputs used above (conserved biophysical features) to identify the largest groups within the human kinome²³; in each case we trained

an SVM with two data sets: the known kinase group served as positives, all other kinase as negatives³⁷. This approach worked best on the cyclin-dependent kinases (CMGC, Table 2). The human CMGC-like group of proteins comprised about 53% of our multi-species true positive cell cycle kinase data set. The tyrosine kinase group (TK, Table 2) was also predicted quite accurately; it comprised <1% of our multi-species cell cycle kinase data set. On the other hand, we failed nicely to identify members of the groups *Other* or *Atypical*. Finally, we also did not really succeed in using CMGC (cyclin-dependent kinases) group alone to identify all cell cycle kinases reaching levels around 27% accuracy at around 57% coverage.

Genome wide classification of human cell cycle kinases

In general, we aimed at adding two types of annotations for human cell cycle kinases. Firstly, we wanted to discover new players, i.e. first time annotations of *domains of unknown function* (DUF) or *hypothetical proteins*. Secondly, we wanted to specifically assign proteins suspected to be kinases to a cellular process. We applied our SVMs to RefSeq Release 22. Depending on our reliability threshold this predicted between 170 (reliable) and 685 (*most daring*) cell cycle S/T kinases. A list of the 170 predictions from the RefSeq sequences is included within the supplementary material. We then investigated to which extent our predictions had added new information: 97 of the 170 reliable predictions (RefSeq sequences) did not have annotations other than a TrEMBL identifier. Of these 97 human RefSeq proteins, 34 align to a non-human TrEMBL protein as their closest homolog. Only 24 of the 97 had >60% sequence identity to ANY protein in our trusted S/T kinase data set; these could have been annotated by homology transfer (expecting about 3–5 mistakes in the transfer for those 24 given the level of similarity, Fig. 2). 43 of the 97 TrEMBL predictions had a PFAM kinase domain and 31 matched a PROSITE kinase pattern, all others could not have been identified by PFAM or PROSITE.

We also PSI-BLASTed the 170 reliable predictions against all human kinases from the kinase.com data set (Table 3). The closest hits ranged from 14% to 100% pairwise sequence identity (PIDE) with a member of the kinase.com data set (the kinase family annotation is provided in the supplementary material). Only four kinase.com groups had representatives with PIDE>90% (Table 3). 130 of the 170 had PIDE≥30% to the kinase.com data set and 84 of those having 90% or greater sequence identity. For those 84, 82% of the predictions are part of only four families STE, CAMK, Other and AGC (Table 3). We find a diversity of genes such as the oxidative-stress response gene OXSR1; CREB1 phosphorylating enzymes KCC1G and TSSK4; CD2L5, a cell-division protein not in our original data set and CO039 an uncharacterized protein with a kinase domain. For all 170 predicted sequences the two top families encountered were: 19% with homology to the STE kinase family and 17% to the CAMK class (Table 3). For the predictions with relation to a member of the CMGC class only 8 of 23 have a percent sequence identity over 42%. Furthermore of the 170 predictions 105 have less than 40% sequence identity to one another using Cd-hit⁷³ and examining non-redundancy via UniqueProt⁴⁹ 36 are sequence unique at an HVAL of 5 while 80 are unique at an HVAL of 15. When examining the 685 predictions made at lower discriminant thresholds, 385 contained only a TrEMBL annotation and 143 align to a non-human protein as their closest homolog. All 42 S/T kinase trusted proteins are found in this set of 685 predictions.

Explicit examples for kinases that we predicted to be implicated in the cell cycle process include the following three. (1) We predicted q17rv0_human (sn112_human) as a cell cycle protein; some limited experimental evidence has linked this protein to the cell cycle⁷⁴. (2) We associated ulk3_human (Entrez GeneID: 25989) with the cell cycle; SWISS-PROT currently annotates this protein only as a “serine-threonine kinase” without specifying the cellular role. (3) The third example may illustrate our goal and our success best: human myotubularin-related protein, mtmr1_human (Entrez GeneID: 8776) is annotated as a phosphatase. Thus, experts

might immediately label our prediction as incorrect. However, this protein has experimentally been implicated with cell cycle arrest in the G1-phase⁷⁵ implying that our prediction could be as right, as it is unexpected given the annotation.

Discussion

Build expert data set for cell cycle by homology-transfer

How can we identify all cell cycle proteins or cell cycle kinases in human? The first solution is simply to begin with a set of all experimentally known cell cycle proteins and to identify all unknown proteins by homology-transfer. Here we did exactly that and established estimates for the accuracy/coverage of such a homology-transfer (Fig. 2, Table 1). As expected, the task is not an easy one in the absence of having very specific structural knowledge for particular families^{76–79}.

The human kinome²³ is built upon a similar idea. However, for our purposes the human kinome classification was too coarse-grained, because our goal was the sub-classification into a cellular process, i.e. the distinction of, e.g. cell cycle Serine/Threonine (S/T) kinases from all other S/T kinases. Databases such as GO⁸⁰ still lack this level of detail (no specific cell cycle kinase class as of yet). Therefore, we had to build our own trusted data set.

Motif-based methods such as PROSITE⁴⁷ can provide a formal annotation for kinases such as the *S/T kinase domain* or the *kinase ATP binding motif* but a more precise biological annotation is still elusive. Classification can be accomplished for particular functions, such as the consensus *CDK phosphorylation site* motif [ST]PX[RK] that may identify kinase substrates²⁸ but does not necessarily translate directly into a biological process annotation such as *cell cycle*. Furthermore, we actually failed when using only the CMGC kinase domain (which contains the PROSITE S/T kinase motif) as our training set to predict cell cycle function.

Predictions for cell cycle and biological relevance

Brunak and co-workers have amply demonstrated that some selective features can be used to predict aspects of protein function directly from sequence⁸¹. Using the framework established for the identification of enzymes, they have shown high coverage (75%) at low false positive rates (15–20%) for the prediction of cell cycle proteins in yeast². For organisms other than yeast, there are currently no proteome-wide predictions available for all proteins involved in cell cycle control. The first stab at collecting what is known about cell cycle in human and other organisms and at developing a method for the prediction of this biological process was our objective.

We examined the role of kinases in the cell cycle process and elucidated several properties of protein sequence, secondary structure and solvent accessibility that contribute to the specificity for this particular biological process. In order to maintain a certain size in their catalytic site clefts, tyrosine kinases and serine-threonine kinases require specific substrate binding residues that reveal themselves as consensus sequences in alignments. In contrast, the distal substrate binding sites vary between the two types of kinases^{45,82}. Although the nucleotide binding site of the kinase family is conserved in sequence and structure, the limited diversity amongst the kinases in this region (for both ATP binding and ATP nonbinding residues) is responsible for mechanistic specificity even enabling successful drug selectivity targeted at this site^{44, 69,70,83–85}. We used evolutionary profiles for very similar proteins (Fig. 1) to select the 30–50% most conserved residues (Fig. 3 and Fig. 6 yellow colored residues) as input for SVMs designed to specifically identify cell cycle kinases.

The correct prediction of amino acid solvent accessibility allows for better modeling of protein secondary structure and kinase inhibitor design therefore research on the improvement of these

prediction methods is constantly ongoing^{86–88}. Nucleotide binding sites among kinases appear in a semi-exposed environment mainly in a cleft between the two kinase lobes^{45,69,85}. We were unable to find in the literature any large-scale analysis for the prediction ability of solvent accessibility for solely ATP and GTP binding residues. We examine our prediction results of solvent accessibility for nucleotide binding residues (Fig. 4). Confirming our ability to correctly classify the solvent accessibility of these specific residues (thereby allowing for their incorporation into our prediction method) is essential to our method and contributes to the overall improvement in prediction ability (Table 1).

In identifying highly conserved residues from functionally similar proteins we sought to mainly capture the residues that contribute to specific functional and structural characteristics such as ligand-binding (Fig. 4). For instance, when we examined the ATP binding sites in human cyclin-dependent kinase 2 (cdk2_human)⁸⁹ our method of using 65% residue conservation PSI-BLAST E-values $\leq 10^{-10}$ retained 13 of 17 ATP binding residues in human cyclin-dependent kinase (Table 4); only Ile10, Phe82, Leu83, Gln131 at frequencies of 32%, 36%, 32% and 56%, respectively were not captured by this threshold (Table 4). For the ATP binding region of the kinase domain of p38 α MAP kinase (mk14_human)^{90,91} our method captured most of the previously identified differences⁶⁹ in the sequence between p38 α and cdk2_human at positions Thr106(26% positional frequency)-Phe80(83%), Leu108(90% positional frequency)-Phe82(36%), Met109(87%)-Leu83(39%), respectively (Fig. 6 green colored residues).

Specifically, the Phe80 residue that acts as a “gatekeeper” of the ATP binding pocket in Cdk2 is highly conserved (at 83% frequency for cutting at $\text{EVAL} < 10^{-10}$). In contrast, the corresponding residues in the p38 α (T105, 26% frequency), Akt2 (M229, 49% frequency) kinases do not exceed our frequency-conservation threshold of 65% to be included as input features (Table 4). Generation of profiles at a less conservative $\text{EVAL} < 10^{-3}$ would not include Phe80 in our feature vector for Cdk2. All these three residues are predicted as buried satisfying our second criterion for feature selection. This diversity at the entry point to the ATP binding pocket among kinases may exactly be the type of signal picked up by our prediction method.

Conversely, the ATP phosphate binding residues (Cdk2 - Gly13, Thr14, Asn132 and Lys33) tend to be more conserved within kinase families^{40,85}. Our feature selection protocol hardly distinguished between these residues (Table 4). The residues surrounding the adenine⁴⁰ are typically hydrophobic in the alignment (e.g. for Val18), with some diversity at Ile10. Backbone interactions with ATP can contribute to the functionality of the binding site if the structure remains conserved, therefore the sequence differences in the site among different kinases can contribute to their specific classification. For instance, backbone amides of the AGC-family α cAMP-dependent protein kinase Ser53, Phe54, Gly55 which contribute to hydrogen bonding with phosphate oxygens correspond to Thr14, Tyr15 and Gly16 of Cdk2 and exhibit diversity among the kinase families (34AYG36 of p38 α ; 162TFG164 of Akt2; 205EFG207 of Csk). The triphosphate group is coordinated by two metals in the DFG motif and in the catalytic loop (Asp127 which may or may not be the proton acceptor in the catalytic mechanism⁹² and Asn132 that stabilizes ATP and the catalytic residues are both highly conserved). Polar groups in the ATP binding region contribute to the integrity of the binding but can vary in sequence among specific kinases. Finally, the CDK2 activating residue Thr160 is only captured above 65% conservation in our profile protocol at an $\text{EVAL} < 10^{-10}$.

Sequence-wise the most conserved portion of ATP binding is in the triphosphoryl binding regions of kinases, while structurally the active conformations exhibit more conservation than the inactive ones⁹³. Other portions of the ATP binding site can show variability in the sequence while still maintaining critical backbone contacts in the active structure conservation. We confirmed many crystallography- and alignment-based studies that showed that a few residues

are highly conserved in the ATP binding site, while other residues provide uniqueness in the binding pocket which can also contribute to specific selectivity and possibly to our prediction ability among cell cycle kinases (Table 4). The differences among kinase families at a stringent definition of conservation contribute to our ability to predict kinases of the cell cycle whereas a more broader definition of conservation does not. In obtaining selected sequence features from the ATP binding site and also the entire protein in our classification protocol these types of differences may pinpoint to the identification of protein function specific to individual types of kinases or to the molecular pathways these kinases are involved in.

We introduced a method that predicts cell cycle S/T kinases at accuracy levels ranging between 70–88% while increasing between 1.4-fold and 1.6-fold in some cases the sensitivity achievable by automatic homology-transfer (Fig. 5). To rule out that only cyclin-dependent kinases contributed to our accurate predictions of cell cycle function in S/T kinase, we used only the CMGC (cyclin-dependent kinase) kinase family³⁷ as our positive training set for which our cell cycle S/T kinase prediction protocol performed significantly worse. Our moderate success in classifying the cell cycle biological process appeared to be based on an implicit (to the machine learning algorithm) recognition of the protein's essential interactions with its substrates. The specific components of the cell cycle allow for this classification; it may not necessarily be the property of other more diverse biological processes.

Conclusions

The ability to predict detailed protein function has significantly progressed in recent years because the identifications of function-specific conserved residues has improved. Here we present another case for the ability to implicitly identify important residues along with their biophysical properties and to use those as a means of improving the predictions of very detailed aspects of function. Our background against which we demonstrated improvement is the best we can do with automated inference by homology, i.e. by similarity to evolutionarily related proteins with experimental annotations. Similar enzymes can act in different cellular pathways, their specificity is often determined by a few key residues responsible for the catalytic mechanism and substrate selectivity. Specific protein kinase function is one such enzymatic activity. Here, we have shown that without knowing the key residue responsible for the specificity, we can use a superset of all somehow conserved and semi-exposed residues that most likely contains most functionally residues to extrapolate a specific protein kinase function to the cell cycle. We applied our methods to un-annotated proteins in human and predicted the cell cycle involvement for 97. These predictions may help researchers shed light on the protein of their interest. More generally, our implicit identification of important residues may allow researchers to distinguish detailed aspects central to a protein's biological process.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Thanks to Dariusz Przybylski (Columbia) and Yanay Ofra (Bar-Ilan University, Israel) for providing preliminary computer programs for accessing certain data used in this study. Thanks to Rajesh Nair (Columbia) for valuable discussions regarding support vector machines and experiment design. Thanks to Drs. Larry Shapiro and Barry Honig (Columbia) for many insights, including detailed explanations from LS about kinase enzymatic mechanisms. Specifically, thanks to all those who deposit their experimental data in public databases and to all those who maintain these databases, especially for all kinase family and sequence data found at <http://kinase.com>. This work was supported by the grant R01-LM07329-01 from the National Library of Medicine (NLM), and by the grant U54-GM074958-01 to the Northeast Structural Genomics consortium (NESG) from the Protein Structure Initiative (PSI) of the National Institutes of Health (NIH).

Abbreviations

AGC	cAMP-kinase/Akt/PKC
ATP	adenosine triphosphate
AUC	Area under ROC curve
CAMK	Ca ²⁺ /calmodulin-dependent kinase
CMGC	cyclin-dependent kinase
CK1	casein kinase 1
DSSP	Definition of the Secondary Structure of Proteins
EC	enzyme nomenclature committee of the international union of biochemistry and molecular biology
EVAL	BLAST/PSI-BLAST expectation value
GTP	guanosine triphosphate
HVAL	HSSP-value, i.e. function correlating pairwise sequence identity and alignment length (Eqn. 1)
ORF	open reading frame
PDB	database of protein structures
PID	percentage of identical residues
PIDE	percentage pairwise sequence identity
PROF	profile-based neural network predictor of protein secondary structure
PROFacc	predictor of solvent accessibility
RGC	receptor guanylate kinase
STE	

MAPK/Ste cascade families

S/T kinase

serine/threonine kinase

SVM

support vector machine

SWISS-PROT

annotated protein sequence database

TK

tyrosine kinase family

TKL

tyrosine kinase-like family

TrEMBL

translation of the EMBL nucleotide database coding DNA to protein sequences

Tyr kinase

tyrosine kinase

References

1. Wrzeszczynski KO, Rost B. Cataloging proteins in cell cycle control. *Methods Mol Biol* 2004;241:219–233. [PubMed: 14970657]
2. de Lichtenberg U, Jensen TS, Jensen LJ, Brunak S. Protein feature based identification of cell cycle regulated proteins in yeast. *J Mol Biol* 2003;329(4):663–674. [PubMed: 12787668]
3. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofra Y. Automatic prediction of protein function. *Cell Mol Life Sci* 2003;60(12):2637–2650. [PubMed: 14685688]
4. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. *Q Rev Biophys* 2003;36(3):307–340. [PubMed: 15029827]
5. Ofra Y, Punta M, Schneider R, Rost B. Beyond annotation transfer by homology: novel protein function prediction methods that can assist drug discovery. *Drug Discovery Today* 2005;10(21):1475–1482. [PubMed: 16243268]
6. Fleming K, Kelley LA, Islam SA, MacCallum RM, Muller A, Pazos F, Sternberg MJ. The proteome: structure, function and evolution. *Philos Trans R Soc Lond B Biol Sci* 2006;361(1467):441–451. [PubMed: 16524832]
7. Mueller M, Martens L, Apweiler R. Annotating the human proteome: beyond establishing a parts list. *Biochim Biophys Acta* 2007;1774(2):175–191. [PubMed: 17223395]
8. Rost, B. Evolution teaches neural networks. In: Clark, JW.; Lindenau, T.; Ristig, ML., editors. *Scientific applications of neural nets*. Heidelberg: Springer; 1999. p. 207–223.
9. Cokol M, Nair R, Rost B. Finding nuclear localization signals. *EMBO Rep* 2000;1(5):411–415. [PubMed: 11258480]
10. Rost B. Enzyme function less conserved than anticipated. *J Mol Biol* 2002;318(2):595–608. [PubMed: 12051862]
11. Wrzeszczynski KO, Rost B. Annotating proteins from Endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes. *Cellular and Molecular Life Sciences* 2004;61(11):1341–1353. [PubMed: 15170512]
12. Mika S, Rost B. Protein–protein interactions more conserved within species than across species. *PLoS Computational Biology* 2006;2(7):e79. [PubMed: 16854211]
13. Nair R, Rost B. Sequence conserved for subcellular localization. *Protein Sci* 2002;11(12):2836–2847. [PubMed: 12441382]

14. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research* 2007;35(11):3823–3835. [PubMed: 17526529]
15. Ofra Y, Rost B. Protein-protein interaction hot spots carved into sequences. *PLoS Computational Biology* 2007;3(7):e119. [PubMed: 17630824]
16. Ofra Y, Rost B. ISIS: Interaction Sites Identified from Sequence. *Bioinformatics* 2007;23(2):e13–16. [PubMed: 17237081]
17. Ofra Y, Rost B. Analysing six types of protein-protein interfaces. *Journal of Molecular Biology* 2003;325:377–387. [PubMed: 12488102]
18. Lichtenberg U, Jensen TS, Brunak S, Bork P, Jensen LJ. Evolution of cell cycle control: same molecular machines, different regulation. *Cell Cycle* 2007;6(15):1819–1825. [PubMed: 17671420]
19. Liu J, Rost B. Domains, motifs, and clusters in the protein universe. *Current Opinion in Chemical Biology* 2003;7:5–11. [PubMed: 12547420]
20. Eyrich VA, Koh IYY, Przybylski D, Graña O, Pazos F, Valencia A, Rost B. CAFASP3 in the spotlight of EVA. *Proteins: Structure, Function, and Bioinformatics* 2003;53(Suppl 6):548–560.
21. Przybylski D, Rost B. Improving fold recognition without folds. *Journal of Molecular Biology* 2004;341:255–269. [PubMed: 15312777]
22. Nair R, Rost B. Mimicking cellular sorting improves prediction of subcellular localization. *Journal of Molecular Biology* 2005;348(1):85–100. [PubMed: 15808855]
23. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science* 2002;298(5600):1912–1934. [PubMed: 12471243]
24. Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999). *European journal of biochemistry/FEBS* 1999;264(2):610–650. [PubMed: 10491110]
25. Hu Z, Mellor J, Wu J, Kanehisa M, Stuart JM, DeLisi C. Towards zoomable multidimensional maps of the cell. *Nat Biotechnol* 2007;25(5):547–554. [PubMed: 17483841]
26. Kobe B, Kampmann T, Forwood JK, Listwan P, Brinkworth RI. Substrate specificity of protein kinases and computational prediction of substrates. *Biochim Biophys Acta* 2005;1754(1–2):200–209. [PubMed: 16172032]
27. Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P. Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* 2006;443(7111):594–597. [PubMed: 17006448]
28. Chang EJ, Begum R, Chait BT, Gaasterland T. Prediction of cyclin-dependent kinase phosphorylation substrates. *PLoS ONE* 2007;2:e656. [PubMed: 17668044]
29. Gauthier NP, Larsen ME, Wernersson R, de Lichtenberg U, Jensen LJ, Brunak S, Jensen TS. Cyclebase.org--a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res* 2008;36(Database issue):D854–859. [PubMed: 17940094]
30. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. *J Mol Biol* 2002;324(1):105–121. [PubMed: 12421562]
31. Petrova NV, Wu CH. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics* 2006;7:312. [PubMed: 16790052]
32. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 2004;20(8):1322–1324. [PubMed: 14871869]
33. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. *J Mol Biol* 1997;271(4):511–523. [PubMed: 9281423]
34. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM. Structure-based activity prediction for an enzyme of unknown function. *Nature* 2007;448(7155):775–779. [PubMed: 17603473]
35. George RA, Spriggs RV, Bartlett GJ, Gutteridge A, MacArthur MW, Porter CT, Al-Lazikani B, Thornton JM, Swindells MB. Effective function annotation through catalytic residue conservation. *Proc Natl Acad Sci U S A* 2005;102(35):12299–12304. [PubMed: 16037208]
36. Lichtarge O, Yao H, Kristensen DM, Madabushi S, Mihalek I. Accurate and scalable identification of functional sites by evolutionary tracing. *J Struct Funct Genomics* 2003;4(2–3):159–166. [PubMed: 14649300]

37. Manning G, Plowman GD, Hunter T, Sudarsanam S. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* 2002;27(10):514–520. [PubMed: 12368087]
38. Hannenhalli SS, Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol* 2000;303(1):61–76. [PubMed: 11021970]
39. Cheek S, Ginalski K, Zhang H, Grishin NV. A comprehensive update of the sequence and structure classification of kinases. *BMC Struct Biol* 2005;5:6. [PubMed: 15771780]
40. Hanks SK, Hunter T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *Faseb J* 1995;9(8):576–596. [PubMed: 7768349]
41. Krupa A, Preethi G, Srinivasan N. Structural modes of stabilization of permissive phosphorylation sites in protein kinases: distinct strategies in Ser/Thr and Tyr kinases. *J Mol Biol* 2004;339(5):1025–1039. [PubMed: 15178245]
42. Taylor SS, Knighton DR, Zheng J, Ten Eyck LF, Sowadski JM. Structural framework for the protein kinase family. *Annu Rev Cell Biol* 1992;8:429–462. [PubMed: 1335745]
43. Taylor SS, Radzio-Andzelm E, Hunter T. How do protein kinases discriminate between serine/threonine and tyrosine? Structural insights from the insulin receptor protein-tyrosine kinase. *Faseb J* 1995;9(13):1255–1266. [PubMed: 7557015]
44. Toledo LM, Lydon NB, Elbaum D. The structure-based design of ATP-site directed protein kinase inhibitors. *Curr Med Chem* 1999;6(9):775–805. [PubMed: 10495352]
45. Ubersax JA, Ferrell JE Jr. Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol* 2007;8(7):530–541. [PubMed: 17585314]
46. Gu J, Bourne PE. Identifying allosteric fluctuation transitions between different protein conformational states as applied to Cyclin Dependent Kinase 2. *BMC Bioinformatics* 2007;8:45. [PubMed: 17286863]
47. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. The PROSITE database. *Nucleic Acids Res* 2006;34(Database issue):D227–230. [PubMed: 16381852]
48. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilboud S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;31(1):365–370. [PubMed: 12520024]
49. Mika S, Rost B. UniqueProt: Creating representative protein sequence sets. *Nucleic Acids Res* 2003;31(13):3789–3791. [PubMed: 12824419]
50. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007;35(Database issue):D61–65. [PubMed: 17130148]
51. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25(17):3389–3402. [PubMed: 9254694]
52. Przybylski D, Rost B. Consensus sequences improve PSI-BLAST through mimicking profile-profile alignments. *Nucleic Acids Res* 2007;35(7):2238–2246. [PubMed: 17369271]
53. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 1991;9(1):56–68. [PubMed: 2017436]
54. Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods in Enzymology* 1996;266:525–539. [PubMed: 8743704]
55. Rost B. Protein secondary structure prediction continues to rise. *Journal of Structural Biology* 2001;134:204–218. [PubMed: 11551180]
56. Rost B, Yachdav G, Liu J. The PredictProtein server. *Nucleic Acids Res* 2004;32(Web Server issue):W321–326. [PubMed: 15215403]
57. Rost B. How to use protein 1D structure predicted by PROFphd. In: Walker, JE., editor. *The Proteomics Protocols Handbook, Methods in Molecular Biology*. Totowa NJ: Humana; 2005. p. 875-901.
58. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins* 1994;20(3):216–226. [PubMed: 7892171]

59. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 2007;35(Database issue):D301–303. [PubMed: 17142228]
60. Ofra Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. *Bioinformatics* 2007;23(13):i347–353. [PubMed: 17646316]
61. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22(12):2577–2637. [PubMed: 6667333]
62. Vapnik, VN. The nature of statistical learning theory. Springer; 1995.
63. Mika S, Rost B. Protein names peeled precisely off free text. *Bioinformatics* 2004;20(Suppl 1):I241–I247. [PubMed: 15262805]
64. Jaakkola T, Diekhans M, Haussler D. A Discriminative Framework for Detecting Remote Protein Homologies. 2000;7:95–114.
65. Nair R, Rost B. LOCnet and LOcTarget: Sub-cellular localization for structural genomics targets. *Nucleic Acids Research* 2004;32(Supplement 2):W517–W521. [PubMed: 15215440]
66. Han LY, Zheng CJ, Xie B, Jia J, Ma XH, Zhu F, Lin HH, Chen X, Chen YZ. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discov Today* 2007;12(7–8):304–313. [PubMed: 17395090]
67. Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNA through support vector machines. *PLoS Genetics* 2006;2(4):e29.10.1371/journal.pgen.0020029 [PubMed: 16683024]
68. Pavlidis P, Wapinski I, Noble WS. Support vector machine classification on the web. *Bioinformatics* 2004;20(4):586–587. [PubMed: 14990457]
69. Noble ME, Endicott JA, Johnson LN. Protein kinase inhibitors: insights into drug design from structure. *Science* 2004;303(5665):1800–1805. [PubMed: 15031492]
70. Fabbro D, Ruetz S, Buchdunger E, Cowan-Jacob SW, Fendrich G, Liebetanz J, Mestan J, O'Reilly T, Traxler P, Chaudhuri B, Fretz H, Zimmermann J, Meyer T, Caravatti G, Furet P, Manley PW. Protein kinases as targets for anticancer agents: from inhibitors to useful drugs. *Pharmacology & therapeutics* 2002;93(2–3):79–98. [PubMed: 12191602]
71. Sprang SR, Coleman DE. Invasion of the nucleotide snatchers: structural insights into the mechanism of G protein GEFs. *Cell* 1998;95(2):155–158. [PubMed: 9790522]
72. Traut TW. The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide-binding sites. *European journal of biochemistry/FEBS* 1994;222(1):9–19. [PubMed: 8200357]
73. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22(13):1658–1659. [PubMed: 16731699]
74. Stephenson A, Huang GY, Nguyen NT, Reuter S, McBride JL, Ruiz JC. snf1lk encodes a protein kinase that may function in cell cycle regulation. *Genomics* 2004;83(6):1105–1115. [PubMed: 15177563]
75. Yoo YD, Cho SM, Kim JS, Chang YS, Ahn CM, Kim HJ. The human myotubularin-related protein suppresses the growth of lung carcinoma cells. *Oncol Rep* 2004;12(3):667–671. [PubMed: 15289854]
76. Korkin D, Davis FP, Sali A. Localization of protein-binding sites within families of proteins. *Protein Sci* 2005;14(9):2350–2360. [PubMed: 16081657]
77. Zhu G, Liu Y, Shaw S. Protein kinase specificity. A strategic collaboration between kinase peptide specificity and substrate recruitment. *Cell Cycle* 2005;4(1):52–56. [PubMed: 15655379]
78. Petrey D, Honig B. Protein structure prediction: inroads to biology. *Mol Cell* 2005;20(6):811–819. [PubMed: 16364908]
79. Endicott JA, Noble ME. Structural principles in cell-cycle control: beyond the CDKs. *Structure* 1998;6(5):535–541. [PubMed: 9634691]
80. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25(1):25–29. [PubMed: 10802651]

81. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S. Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol* 2002;319(5):1257–1265. [PubMed: 12079362]
82. Brinkworth RI, Breinl RA, Kobe B. Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc Natl Acad Sci U S A* 2003;100(1):74–79. [PubMed: 12502784]
83. Breitenlechner C, Gassel M, Hidaka H, Kinzel V, Huber R, Engh RA, Bossemeyer D. Protein kinase A in complex with Rho-kinase inhibitors Y-27632, Fasudil, and H-1152P: structural basis of selectivity. *Structure* 2003;11(12):1595–1607. [PubMed: 14656443]
84. Fabian MA, Biggs WH 3rd, Treiber DK, Atteridge CE, Azimioara MD, Benedetti MG, Carter TA, Ciceri P, Edeen PT, Floyd M, Ford JM, Galvin M, Gerlach JL, Grotzfeld RM, Herrgard S, Insko DE, Insko MA, Lai AG, Lelias JM, Mehta SA, Milanov ZV, Velasco AM, Wodicka LM, Patel HK, Zarrinkar PP, Lockhart DJ. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotechnol* 2005;23(3):329–336. [PubMed: 15711537]
85. Huse M, Kuriyan J. The conformational plasticity of protein kinases. *Cell* 2002;109(3):275–282. [PubMed: 12015977]
86. Pollastri G, Martin AJ, Mooney C, Vullo A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics* 2007;8:201. [PubMed: 17570843]
87. Wang JY, Lee HM, Ahmad S. SVM-Cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine. *Proteins* 2007;68(1):82–91. [PubMed: 17436325]
88. Zarei R, Arab S, Sadeghi M. A method for protein accessibility prediction based on residue types and conformational states. *Comput Biol Chem* 2007;31(5–6):384–388. [PubMed: 17888743]
89. Schulze-Gahmen U, De Bondt HL, Kim SH. High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design. *J Med Chem* 1996;39(23):4540–4546. [PubMed: 8917641]
90. Pargellis C, Tong L, Churchill L, Cirillo PF, Gilmore T, Graham AG, Grob PM, Hickey ER, Moss N, Pav S, Regan J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat Struct Biol* 2002;9(4):268–272. [PubMed: 11896401]
91. Wang Z, Harkins PC, Ulevitch RJ, Han J, Cobb MH, Goldsmith EJ. The structure of mitogen-activated protein kinase p38 at 2.1-Å resolution. *Proc Natl Acad Sci U S A* 1997;94(6):2327–2332. [PubMed: 9122194]
92. De Vivo M, Cavalli A, Carloni P, Recanatini M. Computational study of the phosphoryl transfer catalyzed by a cyclin-dependent kinase. *Chemistry* 2007;13(30):8437–8444. [PubMed: 17636466]
93. Johnson LN, Lowe ED, Noble ME, Owen DJ. The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases. *FEBS Lett* 1998;430(1–2):1–11. [PubMed: 9678585]

Fig. 1. Identifying highly conserved residues

A typical PSI-BLAST alignment is generated for each protein of interest. Homology-transfer of cell cycle annotations is accurate only for very similar proteins. These are selected (shown 80% accuracy threshold). Residue substitution profiles are then compiled for this subset of similar sequences. Only residues in positions that are conserved in 65% of these similar proteins are retained. Positional frequency can greatly diminish at less accurate levels of functional homology (lysine residue in this example) or show initial poor conservation at high levels of functional homology (serine residue in this example).

Fig. 2. Sequence conservation of all trusted cell cycle proteins

We aligned all trusted cell cycle proteins (595) against all true negatives (15,192 proteins combined from all EC classes, all subcellular localizations and proteins annotated as non-enzymes) using PSI-BLAST. Solid lines with filled squares describe cumulative accuracy (percentage of correctly identified cell cycle proteins at given threshold, Eqn. 1); dotted lines with open circles describe cumulative coverage (cell cycle proteins found at threshold/all cell-cycle proteins, Eqn. 2). We measured sequence similarity in three different ways: (A) by the percentage pairwise sequence identity (left graph), (B) the distance from the HSSP-threshold accounting for the length of the alignment (central graph), and (C) by the negative logarithm of the BLAST E-values (note: log to the base of 10) (right graph). For example, the accuracy exceeded 80% for levels > 60% pairwise sequence identity (left), HSSP-distances above 5 (centre), and BLAST expectation values below 10^{-10} (right). At all levels of accuracy ≥ 70 , the HSSP-distance performed best in terms of coverage.

Fig. 3. Fraction of protein vs. conservation

For three different thresholds in deciding which sequences to include in an alignment (different PSI-BLAST thresholds; highest similarity/lowest divergence for top curve, lowest similarity/highest diversity for bottom curve), we plot the fraction of all residues that are conserved (y-axis) vs. the threshold in conservation (x-axis). Our measure for conservation simply is the percentage of proteins that have the same amino acid at a give position in the alignment. For instance, at E-values $<10^{-10}$ about a quarter of all the residues in a protein are conserved between 65% of all proteins (arrow). The bar marks the 65% conservation level that we used as input to our SVMs.

Fig. 4. Confirmation for solvent accessibility prediction among nucleotide binding residues

We present the prediction of solvent accessibility for ATP and GTP binding residues as comparable to DSSP⁶¹ assignments. The x-axes group residues according to their degree of accessible surface from b=buried (black bars), over i=intermediate (gray bars), to e=exposed (white bars); (A) gives values from PROFacc⁵⁸ predictions, (B) shows the accessibility of the same residues from DSSP assignments based on the know structures.

Fig. 5. Cell cycle S/T kinase classification by homology transfer and SVM

We compare the performance in classifying cell cycle S/T kinases by homology transfer (open markers) and by our machine learning approach (SVM, closed markers). Throughout the entire spectrum covered by the ROC curve, our prediction method clearly outperformed the automatic homology transfer. AUC, area under ROC curve values for two types of negative data sets S/T kinase (solid lines) and Tyrosine kinase (dotted lines) favor the SVM method. At a false positive rate of 25% our SVM method outperforms homology transfer 1.4-fold for both negative data sets.

Fig. 6. Conserved structure retained in kinase domains

Two S/T kinase domains depicted with the structure retained by our conserved residue protocol (see Methods and Fig. 1) at 65% positional frequency highlighted in yellow. The top protein is PDB code 1P38⁹¹ the p38 α kinase (mk14_human), the bottom protein is 1HCK⁸⁹ a cyclin-dependent kinase (cdk2_human). Three residues highlighted in green are examples of the differences among the two kinases which were used in the classification protocol between mk14_human and cdk2_human; residues, position and PROFacc prediction are T106(b)-F80(b), L108(b)-F82(b), M109(i)-L83(b), respectively. ATP molecule is in red.

Table 1
Identification of cell cycle Serine-Threonine (S/T) kinases.*

Negative Group	Accuracy all residues	Coverage all residues	Accuracy conserved internal	Coverage conserved internal	Accuracy 60% PIDE	Coverage 60% PIDE
1. <i>all Tyrosine Kinases</i>	0.74 ± .03	0.77 ± .05	0.80 ± .03	0.81 ± .05	0.87	0.38
2. <i>all S/T Kinases</i>	0.63 ± .04	0.65 ± .06	0.70 ± .05	0.62 ± .06	0.86	0.41
3. <i>all proteins</i>	0.87 ± .02	0.69 ± .02	0.88 ± .07	0.69 ± .05	0.54	0.31

* *Negative Group* depicts the type of protein family used as negative examples: *all Tyrosine Kinases* marks the task to identify cell cycle serine-threonine kinases out of all known tyrosine kinases; *all S/T Kinases* marks the task to identify cell cycle serine-threonine kinases out of all known serine-threonine kinases; *all proteins* marks the task to identify cell cycle serine-threonine kinases out of a selected set of all proteins. *Accuracy* and *Coverage* as defined in Eqns. 2 and 3. Columns labeled as all residues present results for using all residues as input to the SVM; columns labeled as *conserved internal* describe using only conserved residues which have predicted buried or intermediate solvent accessibility information (using PROFac) as input; and columns labeled 60% *PIDE* describe using sequence alignment homology transfer at 60% sequence identity for classifying cell cycle serine-threonine kinases.

Table 2

Performance of kinase family classification.*

Human Kinase Group	Accuracy	Coverage
AGC	0.56 ± .05	0.69 ± .05
Atypical	0.43 ± .05	0.88 ± .06
CAMK	0.54 ± .04	0.84 ± .06
CMGC	0.86 ± .05	0.90 ± .06
STE	0.26 ± .06	0.61 ± .08
TK	0.63 ± .05	0.84 ± .06
TKL	0.26 ± .02	0.76 ± .05
Other	0.14 ± .01	0.73 ± .03

* SVM classification accuracy and coverage (Eqns. 2 and 3) for each single individual human kinome family against the remaining human kinome families.

All protein family abbreviations take from ³⁷. **AGC** - *cAMP-kinase/Akt/PKC*, **CAMK** - *Ca²⁺/Calmodulin-dependent kinase*, **CMGC** - *cyclin-dependent kinase*, **STE** - *MAPK/Ste cascade families*, **TK** - *Tyrosine kinase*, **TKL** - *Tyrosine kinase-like*.

Table 3

Human genome predictions relation to kinase family.*

Human Kinase Group	Count	Fraction	Fraction with >90% PIDE
STE	32	0.19	0.21
CAMK	28	0.17	0.24
Other	25	0.15	0.20
CMGC	23	0.14	0.09
AGC	18	0.11	0.17
TK	16	0.09	0.04
TKL	12	0.07	0.04
Atypical	6	0.03	0.00
CK1	6	0.03	0.01
RGC	4	0.02	0.00

* We searched with our top 170 cell cycle predictions with PSI-BLAST against all human kinases in kinase.com. Human Kinase Group: All protein family abbreviations taken from ³⁷: **AGC** - *cAMP-kinase/Akt/PKC*, **CAMK** - *Ca²⁺/Calmodulin-dependent kinase*, **CK1** - *casein kinase 1*, **CMGC** - *cyclin-dependent kinase*, **STE** - *MAPK/Ste cascade families*, **TK** - *Tyrosine kinase*, **TKL** - *Tyrosine kinase-like*, **RGC** - *receptor guanylate kinase*. Count: Number of proteins in the group with the closest PSI-BLAST hit (total over all=170); Fraction: fraction in the group with the closest PSI-BLAST hit (sum=1); Fraction with >90% PIDE: fraction of our predictions in the kinase.com group that have levels of >90% pairwise sequence identity to the closest human representative in the kinase.com group.

Table 4

CDK2 ATP binding residue differences among kinases.*

Residue in Cdk2	Positional Frequency	Residue in p38 α	Positional Frequency	Residue in Akt2	Positional Frequency	Residue in Csk	Positional Frequency
Ile10b	32%	Val30b	35%	Leu158b	65%	Ile201b	41%
Gly13i	100%	Gly33b	100%	Gly161b	99%	Gly204b	97%
Thr14e	81%	Ala34b	87%	Thr162b	17%	Glu205i	27%
Val18b	100%	Val38b	99%	Val166b	99%	Val209b	98%
Ala31b	100%	Ala51b	100%	Ala179b	100%	Ala220b	100%
Lys33i	100%	Lys53b	100%	Lys181i	100%	Lys222i	98%
Val64b	85%	Ile84b	74%	Thr213b	27%	Val249b	68%
Phe80b	83%	Thr106b	26%	Met229b	49%	Thr266b	77%
Glu81b	87%	His107b	8%	Glu230b	59%	Glu267b	100%
Phe82b	36%	Leu108b	90%	Tyr231b	64%	Tyr268b	39%
Leu83b	32%	Met109b	87%	Ala232b	11%	Met269b	84%
Asp86i	99%	Asp112i	95%	Gly235b	100%	Ser273i	57%
Lys129i	100%	Lys152i	99%	Lys277i	98%	Ala316b	77%
Gln131i	56%	Ser154i	66%	Glu279e	56%	Arg318i	74%
Asn132b	100%	Asn155b	100%	Asn280b	98%	Asn319b	95%
Leu134b	98%	Ala157b	20%	Met282b	18%	Leu321b	97%
Asp145b	99%	Asp168b	99%	Asp293b	100%	Asp332b	95%

* The 17 ATP binding site residues of CDK2⁸⁹ are compared to their PSI-BLAST aligned corresponding residue in p38 (CMGC kinase family), Akt2 (AGC kinase family) and Csk (TK kinase family). The 3-letter amino acid code is followed by the amino acid's position in sequence and its predicted solvent accessibility by PROFacc (b=buried, i=intermediate, e=exposed). Positional frequency from our profile generating protocol at PSI-BLAST E-value 10^{-10} (see Methods) of each residue is shown. Amino acids italicized and in gray are completely conserved among all four kinase families. Amino acids with $\geq 65\%$ positional frequency and predicted as buried or intermediate were used in our svm training protocol.