

# Cataloguing proteins in cell cycle control

Kazimierz O. Wrzeszczynski <sup>1</sup>, & Burkhard Rost <sup>1,2,3,\*</sup>

- 1 CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168<sup>th</sup> Street BB217, New York, NY 10032, USA, darek@cubic.bioc.columbia.edulnair@cubic.bioc.columbia.edulrost@columbia.edu
  - 2 Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, NY 10032, USA
  - 3 North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168<sup>th</sup> Street BB217, New York, NY 10032, USA
- \* Corresponding author: rost@columbia.edu, <http://cubic.bioc.columbia.edu/>  
Tel: +1-212-305-3773, fax: +1-212-305-7932

**Running Title:** Cataloguing proteins in cell cycle control  
**Document statistics:** Abstract 202, Text = 2415 words, 70 references, 2 figures; 3 tables  
**Book Title:** Cell cycle checkpoint control protocols  
**Book Editor:** Howard Lieberman  
**Book Publisher:** Humana Press  
**Journal:** Methods in Molecular Biology  
**Submitted:** July 1, 2002

# Cataloguing proteins in cell cycle control

Kazimierz O Wrzeszczynski<sup>1</sup> & Burkhard Rost<sup>1,2,3\*</sup>

## Abstract

Bioinformatics makes a number of methods available that can also be used to identify cell cycle related proteins. Nevertheless, few tools are specifically designed to cope with cell cycle proteins. In fact, a vast amount of data is currently scattered among many databases. Here, we present a first detailed analysis of known cell cycle proteins. We combined databases mining and literature searches with an evaluation of evolutionary conservation. The objective was to identify cell cycle control proteins in various proteomes. In total we found 595 experimentally annotated cell cycle control proteins; these clustered into 113 distinct structural families. We noticed that neither simple values for pairwise sequence identity nor expectation values taken from popular PSI-BLAST alignments allow an error-free inference of involvement in cell cycle control by sequence similarity. However, when we also considered alignment length we could find thresholds for reliable inference of cell cycle proteins. Applying these safe thresholds to the six entirely sequenced organisms (human, mouse, fly, worm, arabidopsis and yeast), we could identify 463 un-annotated proteins likely to be involved in cell cycle control. Slightly lower levels of accuracy extended the count to approximately 500-1300 additional proteins, which may be candidates for involvement in cell cycle control process.

---

\* Corresponding author: [rost@columbia.edu](mailto:rost@columbia.edu)

1 CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168<sup>th</sup> Street BB217, New York, NY 10032, USA,

2 Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, NY 10032, USA

3 North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168<sup>th</sup> Street BB217, New York, NY 10032, USA

## 1. Introduction □

*No direct path from sequence similarity to 'biological' similarity.* How can bioinformatics tools help to identify particular types of proteins? In general, the answer depends on the type of protein. Alignment methods can identify similarities between two proteins. However, while database search tools are optimised to finding the best possible superposition between two proteins, they fail in answering questions such as: Does the query protein Q perform the same function as the protein in the database H for which we have some experimental data about function? In fact, alignment methods typically provide some statistical score evaluating the probability that the similarity between Q and H happened by chance (1, 3). The precise function relating such a statistical score for sequence similarity to the actual 'biological' similarity of two proteins, i.e. similarity in terms of their three-dimensional (3D) structure and/or function depends on the problem. For example, if the PSI-BLAST expectation value for the similarity between Q and H is below  $10^{-5}$ , then this typically implies that H and Q have similar local 3D structure (6). However, less than 70% of all pairs of enzymes that have this level of sequence similarity have exactly the same enzymatic activity (7) and over 90% of all pairs with so similar sequences are observed in the same sub-cellular compartment (8). Establishing these estimates typically requires solving three different tasks: (1) define biological similarity (3D, enzyme activity, sub-cellular localization), (2) build unbiased data sets of experimentally reliable information, (3) and establish thresholds that relate sequence to biological similarity. These steps have been completed for a variety of biological features such as structure (6, 9, 10, 11, 12), enzymatic activity (7, 13, 14, 15, 16, 17), active sites (14), binding sites (14), functional keywords (14), functional classes (14, 16), sub-cellular localization (8, 18). However, there is no way to infer from these results at which level of sequence similarity we can conclude that two homologous proteins play the same role in processes such as cell cycle control.

The field of proteomics has evolved into various levels of biological and computational techniques that identify and

---

□ **Abbreviations used:** **3D structure**, three-dimensional coordinates of protein structure; **BLAST**, fast sequence alignment method (1); **PDB**, Protein Data Bank of experimentally determined 3D structures of proteins (2); **PSI-BLAST**, position specific iterated database search (3); **Q**, query protein, i.e. protein used to search the database for homologues, **SRS**, Sequence Retrieval System, i.e. the most general portal allowing to simultaneously access most existing data bases (4); **SWISS-PROT**, data base of protein sequences (5); **TrEMBL**, translation of the EMBL-nucleotide database coding DNA to protein sequences (5).

classify proteins in the context of entire genomes and proteomes. These techniques include a broad spectrum of approaches; from detailed literature searches (19, 20) or text analysis of database annotations (21), database mining (20, 22, 23, 24, 25, 26, 27, 28), multiple sequence alignments (3, 29, 30, 31, 32, 33), protein family clustering (34, 35, 36, 37, 38, 39), methods predicting aspects of protein function and structure (40, 41, 42, 43, 44, 45, 46) and computational modelling of the cell cycle (47) to gene microarray or 'chip' expression techniques (48, 49, 50, 51, 52), yeast two-hybrid systems (53, 54) and recently mass spectroscopy of protein complexes (55, 56). The process of unifying these techniques from an assortment of cataloguing tools into a more eloquent analysis of the cell cycle and specifically cell cycle control proteins is only beginning to take shape. Here, we present a first step for this process using database mining and literature searches to evaluate the current status of cell cycle control proteins present in various databases, combined with sequence alignment evaluation to identify cell cycle control proteins in various proteomes. We began by archiving proteins known to be involved in cell cycle control through database and literature searches. Then, we established levels of sequence similarity that imply similarity in function. Finally, we attempted identifying cell cycle control proteins through homology in entirely sequenced eukaryotic proteomes.

## 2. Materials

### 2.1. Public databases

Curated, well-formatted and annotated databases comprise one of the most important resources for bioinformatics. A few public databases contain information about cell cycle proteins (Table 1); from these we built a resource that identifies the general register of cell cycle information currently available. To create this repository, we collected about 3811 records from MEDLINE (57). Using SRS (4), we retrieved about 364 proteins from SWISS-PROT (5), and 98 proteins of known structure from PDB (2). Only seven of these 98 were classified as 'cell cycle control' proteins. A closer inspection of the SWISS-PROT dataset revealed 534 proteins with the keyword 'cell cycle', and 940 with the keyword 'cell division'. ProtoNet (36, 58) is a tool that clusters all proteins from SWISS-PROT into somehow related families. ProtoNet identified 1476 clusters with a total of 512 proteins for the SWISS-PROT keyword 'cell cycle' and 887 proteins in 1983 clusters with the keyword 'cell division'. The obvious next task was to peel out a catalogue of unique families of proteins related to cell cycle (Methods).

### 2.2. Sources of sequences for entire proteomes

All human sequences were extracted from SWISS-PROT and TrEMBL (5). We retrieved all other proteome sequences from the respective public sites: *Drosophila melanogaster*: <http://www.fruitfly.org/>, *Caenorhabditis elegans*: <ftp://ncbi.nlm.nih.gov/genbank/genomes/>, *Saccharomyces cerevisiae* from the yeast genome directory (59), *Arabidopsis Thaliana*: <http://www.arabidopsis.org/>, and *Mus Musculus*: <http://www.ensembl.org>.

## 3. Methods

### 3.1. Cell cycle and cell cycle control proteins in public databases

*Keyword search in SWISS-PROT.* First, we searched for proteins of trusted experimental information about cell cycle control in SWISS-PROT. Most proteins retrieved thus control the g1/s and g2/m transitions, or are related to the m and s phases. In total, we found 361 proteins (Table 2) that were distributed amongst various species. Next, we clustered these proteins into families.

*Sequence-unique data sets.* 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 that no pair in the set had more than 33% identical residues over more than 100 residues aligned (HSSP-threshold of 0 (6)). 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 condition. This reduced the entire set of 361 to 42 unique proteins or protein families.

*Extending simple keyword-based search.* 42 unique proteins did not suffice to develop any statistical criteria for determining levels of significant sequence similarity and also implying similarity in the cell cycle process. We expanded our original data set by including searches for other cell cycle controlling factors such as ubiquitin, and those in the ras super-family, plus other proteins annotated for cell division control. This extensive search for cell cycle control proteins increased the list to a total of 595 proteins; 97 of these had multiple, conflicting annotations (Table 2); 113 were sequence-unique, i.e. we increased the numbers of families from 42 to 113 through the extended keyword-based search. The entire dataset of cell cycle control proteins is in the preparation of being made available online at the CUBIC website: [cubic.bioc.columbia.edu](http://cubic.bioc.columbia.edu).

**Table 1: Public resources for cell cycle proteins <sup>a</sup>****Databases**

The Suiseki Information Extraction System	<a href="http://www.pdg.cnb.uam.es/suiseki/">www.pdg.cnb.uam.es/suiseki/</a>
Yeast Cell Cycle Analysis Project	<a href="http://www.pdg.cnb.uam.es/suiseki/system/Start_cellCycle_new.html">www.pdg.cnb.uam.es/suiseki/system/Start_cellCycle_new.html</a>
SCPD: Promoter Database of <i>Saccharomyces cerevisiae</i>	<a href="http://genome-www.stanford.edu/cellcycle/data/rawdata/">genome-www.stanford.edu/cellcycle/data/rawdata/</a>
Mouse Genome Informatics	<a href="http://cgsigma.cshl.org/jian/">cgsigma.cshl.org/jian/</a>
The Interactive Fly - Cell Cycle in <i>Drosophila</i>	<a href="http://www.informatics.jax.org">www.informatics.jax.org</a>
Transfac & Transpath	<a href="http://sdb.bio.purdue.edu/fly/aimain/aadevinx.htm">sdb.bio.purdue.edu/fly/aimain/aadevinx.htm</a>
Mitosis World	<a href="http://transfac.gbf.de/TRANSFAC/">transfac.gbf.de/TRANSFAC/</a>
TRRD - Transcription Regulatory Regions Database	<a href="http://www.bio.unc.edu/faculty/salmon/lab/mitosis/mitosis.html">www.bio.unc.edu/faculty/salmon/lab/mitosis/mitosis.html</a>
The Ubiquitin System for Protein Modification and Degradation	<a href="http://www.bionet.nsc.ru/trrd/">www.bionet.nsc.ru/trrd/</a>
KEGG: Kyoto Encyclopedia of Genes and Genomes	<a href="http://http://wwwmgs.bionet.nsc.ru/mgs/papers/ke1_ov/celcyc/">http://wwwmgs.bionet.nsc.ru/mgs/papers/ke1_ov/celcyc/</a>
The p53 web site	<a href="http://www.nottingham.ac.uk/biochemcourses/students/ub/ubindex.html">www.nottingham.ac.uk/biochemcourses/students/ub/ubindex.html</a>
The Kinesin Home Page	<a href="http://www.genome.ad.jp/kegg/">www.genome.ad.jp/kegg/</a>
The Database for Interacting Proteins	<a href="http://www.genome.ad.jp/kegg/pathway/hsa/hsa04110.html">www.genome.ad.jp/kegg/pathway/hsa/hsa04110.html</a>
The Forsburg Lab pombe Pages	<a href="http://p53.curie.fr/">p53.curie.fr/</a>
Protonet - Automatic Hierarchical Classification of Proteins	<a href="http://www.proweb.org/kinesin/">www.proweb.org/kinesin/</a>
MIPS - Comprehensive Yeast Genome Database	<a href="http://www.proweb.org/kinesin//KinesinTree.html">www.proweb.org/kinesin//KinesinTree.html</a>
Protein Information Resource	<a href="http://dip.doe-mbi.ucla.edu/">dip.doe-mbi.ucla.edu/</a>
PDB: database of protein structures	<a href="http://pingu.salk.edu/~forsburg/lab.html">pingu.salk.edu/~forsburg/lab.html</a>
SWISS-PROT (annotated proteins)	<a href="http://www.protonet.cs.huji.ac.il/protonet/index.php">www.protonet.cs.huji.ac.il/protonet/index.php</a>

**Tools**

PSI-BLAST (database search)	<a href="http://www.ncbi.nlm.nih.gov/BLAST/">www.ncbi.nlm.nih.gov/BLAST/</a>
Predictions of post-translational modifications	<a href="http://www.cbs.dtu.dk/services/">www.cbs.dtu.dk/services/</a>
PredictProtein (sequence analysis + structure prediction)	<a href="http://cubic.bioc.columbia.edu/predictprotein">cubic.bioc.columbia.edu/predictprotein</a>
META-PP (interface to variety of tools)	<a href="http://cubic.bioc.columbia.edu/predictprotein/submit_meta.html">cubic.bioc.columbia.edu/predictprotein/submit_meta.html</a>
ExPasy (tools, databases, links)	<a href="http://www.expasy.ch/">www.expasy.ch/</a>
WWW links for molecular biology	<a href="http://cubic.bioc.columbia.edu/doc/links_index.html">cubic.bioc.columbia.edu/doc/links_index.html</a>

<sup>a</sup> Note 1: we dropped the string 'http://' from the URL, e.g. to access KEGG you may have to type 'http://www.genome.ad.jp/kegg' in some browsers

Note 2: we will make all our data along with a novel cell cycle specific database available through our website [cubic.bioc.columbia.edu](http://cubic.bioc.columbia.edu)

**Table 2: Numbers of cell cycle control proteins found in SWISS-PROT <sup>1</sup>**

<i>Species</i>	<i>cell cycle control</i>	<i>g1/s</i>	<i>g2/m</i>	<i>m phase</i>	<i>s phase</i>	<i>other</i>	<i>multiple</i>
<i>Eukaryotes</i>	582	135	86	66	156	229	90
<i>Homo sapiens</i>	99	28	11	23	41	24	28
<i>Mus musculus</i>	68	25	8	10	30	18	23
<i>Drosophila melanogaster</i>	15	5	3	2	4	3	2
<i>Caenorhabditis elegans</i>	10	1	4	1	2	2	0
<i>Arabidopsis thaliana</i>	5	0	1	0	0	4	0
<i>Saccharomyces cerevisiae</i>	87	20	11	5	19	46	14

<sup>1</sup> Eukaryotic proteins presented, the remainder of proteins in the set of 595 cell cycle proteins are involved in the prokaryotic cell cycle process.

### 3.2. Cell cycle control protein identification through sequence similarity

*Establishing threshold for significant sequence similarity.* 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. Obviously, such thresholds have to find a balance between accuracy and coverage, in other words, we have to navigate between the Skylla of 'high selectivity/low sensitivity', i.e. finding very few homologues all of which are right, and the Charibdis of 'low selectivity/high sensitivity', i.e. finding many putative homologues, most of which are wrong. Cumulative accuracy and coverage were calculated as:

$$\text{Cumulative Accuracy} = 100 \cdot \frac{\text{number of true pairs found above threshold}}{\text{number of all pairs above threshold}} \quad (1)$$

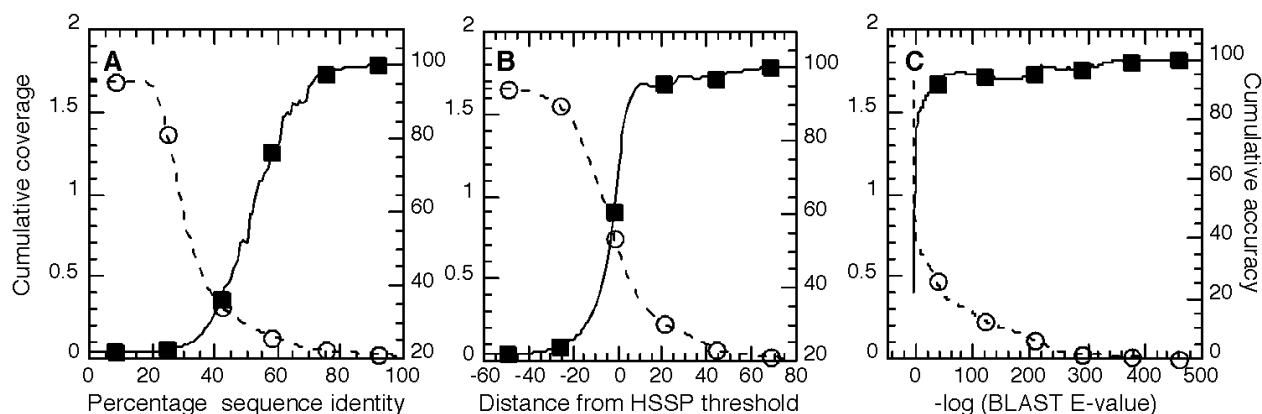
$$\text{Cumulative Coverage} = 100 \cdot \frac{\text{number of true pairs found above threshold}}{\text{number of all true pairs}} \quad (2)$$

with the thresholds for sequence similarity specified below.

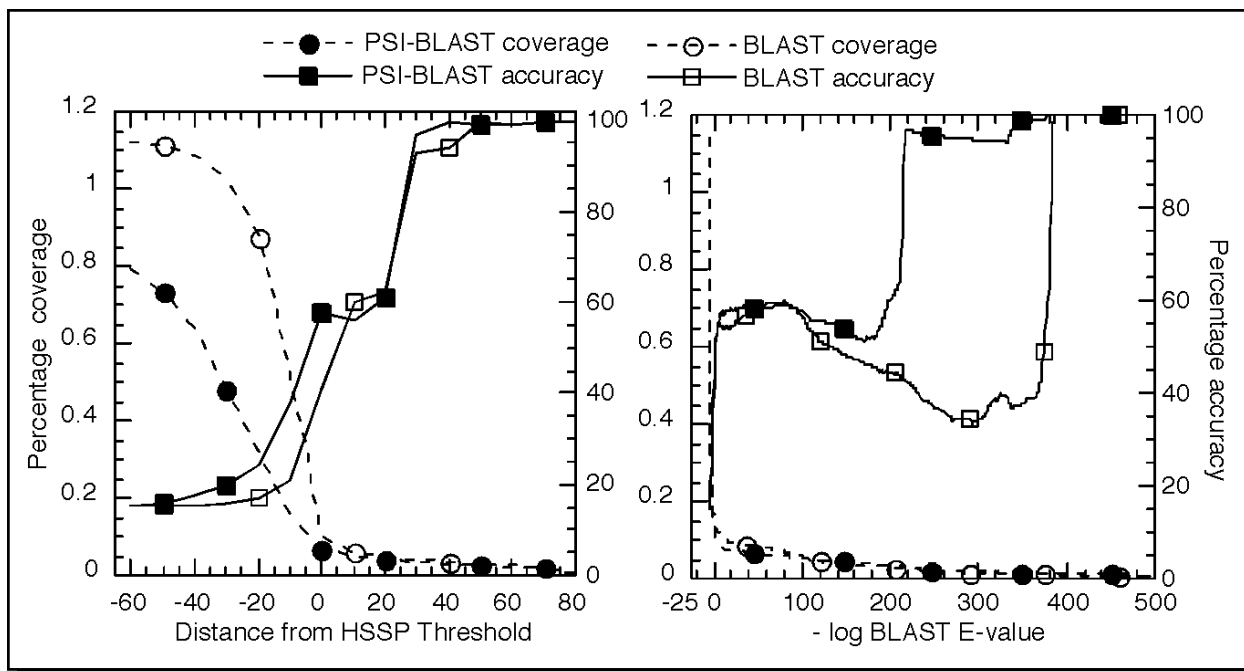
*Aligning proteins.* We generated alignments for all sequences from the cell cycle unique dataset (595)

against a set of non-nuclear (but including cytoplasmic) proteins of known function other than those functions in cell cycle control (total of 6728 proteins) using pairwise BLAST (1). To refine the analysis, we also generated PSI-BLAST profiles using a filtered version of all currently known sequences with three iterations (60). These profiles were then aligned against our 'cell cycle control plus all other proteins' dataset. Sequence similarity was defined by percentage identity, BLAST E-values, and the distance from the HSSP-threshold which relates percentage sequence identity to alignment length thus accounting for the fact that 80% pairwise identity is not significant when achieved over a stretch of 15 consecutive residues, however, it is highly informative when achieved over entire proteins (61).

*Accuracy and coverage of inferring cell cycle role by homology.* When we aligned all trusted cell cycle proteins (595) against all true negatives (6116 non-cell cycle proteins), we found that at HSSP-distances of 15 (corresponding to 48% pairwise sequence identity for more than 100 aligned residues), we could seemingly infer the role in the cell cycle at an accuracy of 95% (Fig. 1). However, when using the unbiased, sequence-unique subset of 113 cell cycle proteins to evaluate accuracy, we found levels of only 60% accuracy. In order to reach a level of 95% accuracy, we had to increase the HSSP-distance from 15 to 40 (Fig. 2), i.e. have to require over



**Fig. 1: Sequence conservation of all trusted cell cycle control proteins.** We aligned all trusted cell cycle proteins (595) against all true negatives (6116 non-cell cycle proteins) using 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 3 (centre), and BLAST expectation values below  $10^{-12}$  (right). At all levels of accuracy  $\geq 80$ , the HSSP-distance performed best in terms of coverage. Note that these estimates were based on large data sets, however, they constituted over-estimates, since the bias in the data sets was not removed.



**Fig. 2: Estimating accuracy and coverage for BLAST and PSI-BLAST.** In order to correctly estimate the likely accuracy and coverage, we had to remove the bias from our initial data sets by aligning the subset of 113 sequence unique trusted cell-cycle proteins against all trusted cell-cycle proteins and against all true negatives. For this, we compared the performance of pairwise BLAST (open symbols) to that of PSI-BLAST (filled symbols). Accuracy (solid lines) and coverage (dashed lines with circles) were as in Fig. 1. In general, PSI-BLAST clearly outperformed BLAST. For example, at HSSP-distances  $> 40$  the accuracy of PSI-BLAST searches was above 95%. Note that these estimates were sufficiently lower than those that would have been obtained using the biased data (Fig. 1). Using only the E-values taken from PSI-BLAST and BLAST alignments required very high cut-off thresholds: even at levels of  $10^{10}$ , implying that only one in ten million hits occurred by chance, less than 70% of the inferences were correct. The residual problem with the data resulted from the small set sizes (rigged curves).

70% pairwise sequence identity). Replacing the HSSP-distance by the expectation values from BLAST or PSI-BLAST (E-values) did not yield a more accurate distinction between true and false positives. This finding confirmed our previous results on establishing thresholds for sequence similarity implying similarity in 3D structure and sub-cellular localisation (7, 8).

*Identifying cell cycle control proteins from entirely sequenced proteomes.* We used a variety of thresholds for inferring the role of cell cycle control proteins by homology as to confer the annotations about these roles from our trusted data set to homologues in entirely sequenced eukaryotes. In particular, we scanned the proteomes of human (*Homo sapiens*), mouse (*Mus musculus*), fly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*), weed (*Arabidopsis thaliana*), and yeast (*Saccharomyces cerevisiae*). At levels of around 95% accuracy, we could extend the number of proteins known to be involved in cell cycle control from 284 for the six completely sequenced organisms to about

747 (Table 3). Our analysis also pulled out about 500-1300 additional proteins (difference between columns D=40 and D=25 and D=15 in Table 3) that may constitute candidates for unknown cell-cycle control proteins. On the other extreme end, our data illustrated that over 10000 proteins in any of these six proteomes have similar 3D structures to one of the known cell-cycle proteins. Supposedly most of these are not related to cell-cycle control, illustrating the variety of functions that can be adopted by proteins of similar structure.

#### 4. Notes

##### 4.1. Limits of inferring function through homology

Everyday biologists are searching with their protein Q of interest by standard alignment methods to uncover putative homologies to their protein. Due to large-scale sequencing efforts, these database searches retrieve more

and more often proteins without any annotation other than 'hypothetical protein'. To initiate hypotheses about function such results are obviously not very informative. More difficult are the 'helpful' cases when a protein with experimental annotation about function H is similar to Q. The number of pitfalls that can lead to incorrect hypotheses based on database searches are manifold (14, 28, 62, 63, 64, 65). Nevertheless, an increasing number of publications in modern biology is based on some beneficial hints obtained from database searches. How can we separate the chaff from the wheat? Certainly, it is a *sine qua non* to establish thoroughly evaluated, statistically significant estimates for which level of sequence similarity implies what (7, 13, 14, 16, 17, 66). In the context of cell cycle proteins, our approach aims at identifying commonalities in the evolutionary conservation of a selected group of functions. On the one hand, it appears evident that all proteins involved in cell cycle and cell cycle control have common evolutionary constraints. If true we can infer the involvement of a protein in the cell cycle process based on sequence similarity. On the other hand, we may suspect that two kinases such as pyruvate dehydrogenase kinase and Cdk1 are more similar than the two cell cycle proteins Cdk1 kinase and the E2F transcription factor. If true, we have to define all types of function related to cell cycle and have to establish thresholds for each functional type; in other words, our inference of cell cycle roles based on homology is rather limited. Arguably reality falls between these two extremes. Therefore, our ability to discover new proteins in cell cycle control through homology works to some extent, but is rather restricted.

#### 4.2. Other tools targeting cell cycle proteins

Jones & Sgouros (67) studied cohesion complex proteins through sequence motifs and database searches. They used PSI-BLAST to identify all homologues of the SMC (Structural Maintenance of Chromosomes) and the SCC (Sister-Chromatid Cohesion) proteins from yeast (Smc1, Smc3, Scc1, Scc2, Scc3, and Scc4), as well as four proteins interacting with cohesion proteins (Trf4, Prp11, Tid3, Esp1). Next, the authors aligned the putative homologues identified by PSI-BLAST using the dynamic programming based method ClustalX (32, 33), and constructed putative evolutionary trees from these ClustalX alignments using the program PHYLIP (68). Finally, the study identified possible binding partners from the complete two-hybrid screens available through the Yeast Proteome Database and putative sequence motifs through the program Teiresias (69). The study resulted in the establishment of five families of SMC proteins, a cohesion interaction network of 17 proteins and the identification of possible common sequence motifs for binding and a kinase active site.

Kel and colleagues (70) combined experimental and theoretical techniques in a comprehensive study identifying the 5' regulatory regions of cell-cycle related genes. First, the group developed a program that identifies context-specific binding sites for the E2F transcription factors. All these sites were identified in entirely sequenced genomes with the aim to identify new genes that play a role in controlling cell-proliferation, differentiation, and apoptosis. Finally, the predictions were verified by chromatin immunoprecipitation assays. The study resulted in a total of 313 new potential E2F targets found, 8 of which were verified through the *in vivo* experimentation.

Blaschke & Valencia (19) developed a text analysis system (SUISEKI: System for Information Extraction on Interactions) that automatically identifies cell-cycle related protein-protein interactions from scientific literature, i.e. from MEDLINE abstracts. At the heart of the system, text searches are defined into frames that capture the various language constructs used to convey protein interactions. The authors selected 5,283 abstracts

**Table 3: Cell cycle control proteins predicted by homology in entire proteomes<sup>1</sup>**

Proteome	Known cell cycle control proteins <sup>2</sup>	Predicted cell cycle control proteins			
		D=0 (55%)	D=15 (65%)	D=25 (90%)	D=40 (95%)
<i>Homo sapiens</i>	99	3073	782	476	299
<i>Mus musculus</i>	68	3162	574	310	203
<i>Drosophila melanogaster</i>	15	970	181	96	50
<i>Caenorhabditis elegans</i>	10	1005	185	87	32
<i>Arabidopsis thaliana</i>	5	1888	303	148	63
<i>Saccharomyces cerevisiae</i>	87	513	148	119	100
<b>Sum</b>	<b>284</b>	<b>10611</b>	<b>2173</b>	<b>1236</b>	<b>747</b>

- 1 Distance from HSSP-Threshold chosen as seen in Fig. 2 for various levels of percent accuracy using the PSI-BLAST curve. Levels of accuracy are estimated according to Fig. 2, e.g. at a threshold of D=40 more than 95% of the proteins for which we infer the involvement in cell cycle control by homology are supposedly correctly inferred.
- 2 The number of previously known annotated cell cycle control proteins represented in each specific proteome as used in our trusted data set is given for comparison.

that included the word “cell-cycle”, the system detected 6,778 protein interactions from all of the abstracts, resulting finally in 4,657 distinct interactions from a total of 1,471 abstracts. The data is currently available at [www.pdg.cnb.uam.es/suiseki/](http://www.pdg.cnb.uam.es/suiseki/).

## 5. Acknowledgements

Thanks to Jinfeng Liu (Columbia) for computer assistance and the collection of genome data sets; to Jinfeng Liu, Dariusz Przybylski (Columbia), and Rajesh

Nair (Columbia) for providing preliminary information and programs. Particular thanks to Volker Eyrich (Columbia) for programming and maintaining most of the immensely valuable software that runs the EVA and META-PredictProtein servers! The work of JL and BR was supported by the grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

## 6. References

1. Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods in Enzymology* 266, 460-480.
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* 28, 235-42.
3. Altschul, S., Madden, T., Shaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Research* 25, 3389-3402.
4. Etzold, T., Ulyanov, A. & Argos, P. (1996). SRS: Information retrieval system for molecular biology data banks. *Methods in Enzymology* 266, 114-128.
5. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28, 45-8.
6. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng* 12, 85-94.
7. Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of Molecular Biology* 318, 595-608.
8. Nair, R. & Rost, B. (2002). Sub-cellular localisation surprisingly conserved in sequence. *Protein Science*, submitted.
9. Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Genetics* 9, 56-68.
10. Abagyan, R. A. & Batalov, S. (1997). Do aligned sequences share the same fold? *Journal of Molecular Biology* 273, 355-368.
11. Alexandrov, N. N. & Soloveyev, V. V. (1998). *HICCS' 98: Pacific Symposium on Biocomputing' 98, Maui, Hawaii, U.S.A.*
12. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences* 95, 6073-6078.
13. Shah, I. & Hunter, L. (1997). *Fifth International Conference on Intelligent Systems for Molecular Biology, Halkidiki, Greece.*
14. Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Structure, Function, and Genetics* 41, 98-107.
15. Jaroszewski, L., Rychlewski, L. & Godzik, A. (2000). Improving the quality of twilight-zone alignments. *Protein Science* 9, 1487-1496.
16. Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology* 297, 233-249.
17. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology* 307, 1113-1143.
18. Wrzeszczynski, K. O. & Rost, B. (2002). Retention signals for Endoplasmic reticulum and Golgi apparatus motifs inaccurate. *Proteins: Structure, Function, and Genetics*, in preparation.
19. Blaschke, C. & Valencia, A. (2001). The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform Ser Workshop Genome Inform* 12, 123-34.
20. Valencia, A. (2002). Search and retrieve: Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data? *EMBO Reports* 3, 396-400.
21. Nair, R. & Rost, B. (2002). Inferring sub-cellular localisation through automated lexical analysis. *Bioinformatics*, in press.
22. Walker, D. R. & Koonin, E. V. (1997). *Fifth International Conference on Intelligent Systems for Molecular Biology, Halkidiki, Greece.*
23. Schmitt, A. O., Specht, T., Beckmann, G., Dahl, E., Pilarsky, C. P., Hinzmann, B. & Rosenthal, A.

- (1999). Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Research* 27, 4251-60.
24. Andrade, M. A. & Bork, P. (2000). Automated extraction of information in molecular biology. *FEBS Lett* 476, 12-7.
  25. Gaasterland, T., Sczyrba, A., Thomas, E., Aytekin-Kurban, G., Gordon, P. & Sensen, C. W. (2000). MAGPIE/EGRET annotation of the 2.9-Mb *Drosophila melanogaster* Adh region. *Genome Res* 10, 502-510.
  26. Galperin, M. Y. & Koonin, E. V. (2000). Who's your neighbor? New computational approaches for functional genomics. *Nature Biotechnology* 18, 609-613.
  27. Gaasterland, T. & Oprea, M. (2001). Whole-genome analysis: annotations and updates. *Current Opinion in Structural Biology* 11, 377-381.
  28. Koonin, E. V. (2001). Computational genomics. *Curr Biol* 11, R155-8.
  29. Smith, T. F., Waterman, M. S. & Burks, C. (1985). The statistical distribution of nucleic acid similarities. *Nucl. Acids Res.* 13, 645-656.
  30. Higgins, D. G., Thompson, J. D. & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology* 266, 383-402.
  31. Pearson, W. R. (1996). Effective protein sequence comparison. *Methods in Enzymology* 266, 227-258.
  32. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends in Biochemical Sciences* 23, 403-405.
  33. Higgins, D. G. & Taylor, W. R. (2000). Multiple sequence alignment. *Methods Mol Biol* 143, 1-18.
  34. Enright, A. J. & Ouzounis, C. A. (2000). GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* 16, 451-457.
  35. Gerstein, M. & Jansen, R. (2000). The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? *Curr Opin Struct Biol* 10, 574-584.
  36. Linial, M. & Yona, G. (2000). Methodologies for target selection in structural genomics. *Progress in Biophysics and Molecular Biology* 73, 297-320.
  37. Heger, A. & Holm, L. (2001). Picasso: generating a covering set of protein family profiles. *Bioinformatics* 17, 272-279.
  38. Rehmsmeier, M. & Vingron, M. (2001). Phylogenetic information improves homology detection. *Proteins: Structure, Function, and Genetics* 45, 360-371.
  39. Liu, J. & Rost, B. (2002). Target space for structural genomics revisited. *Bioinformatics*, in press.
  40. Jones, D. T. (1997). Progress in protein structure prediction. *Current Opinion in Structural Biology* 7, 377-387.
  41. Rost, B. & Sander, C. (2000). Third generation prediction of secondary structure. *Methods in Molecular Biology* 143, 71-95.
  42. Thornton, J. W. & DeSalle, R. (2000). Gene family evolution and homology: genomics meets phylogenetics. *Annu Rev Genomics Hum Genet* 1, 41-73.
  43. Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science* 294, 93-96.
  44. Pawlowski, K., Rychlewski, L., Zhang, B. & Godzik, A. (2001). Fold predictions for bacterial genomes. *Journal of Structural Biology* 134, 219-231.
  45. Rost, B. (2001). Protein secondary structure prediction continues to rise. *Journal of Structural Biology* 134, 204-218.
  46. Rost, B. (2002). Did evolution leap to create the protein universe? *Current Opinion in Structural Biology* 12, 409-416.
  47. Tyson, J. J. & Novak, B. (2001). Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *J Theor Biol* 210, 249-63.
  48. Gaasterland, T. & Bekiranov, S. (2000). Making the most of microarray data. *Nature Genetics* 24, 204-206.
  49. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. & Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* 29, 365-371.
  50. Cho, R. J., Huang, M., Campbell, M. J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S. J., Davis, R. W. & Lockhart, D. J. (2001). Transcriptional regulation and function during the human cell cycle. *Nat Genet* 27, 48-54.
  51. Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., Kaloper, M., Weng, S., Jin, H., Ball, C. A., Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. & Cherry, J. M. (2001). The Stanford Microarray Database. *Nucleic Acids Res* 29, 152-5.
  52. Shedden, K. & Cooper, S. (2002). Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc Natl Acad Sci U S A* 99, 4379-84.

53. Cagney, G., Uetz, P. & Fields, S. (2000). High-throughput screening for protein-protein interactions using two-hybrid assay. *Methods Enzymol* 328, 3-14.
54. Tucker, C. L., Gera, J. F. & Uetz, P. (2001). Towards an understanding of complex protein networks. *Trends Cell Biol* 11, 102-6.
55. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edlmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141-7.
56. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D. & Tyers, M. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180-3.
57. Airozo, D., Allard, R., Brylawski, B., Canese, K., Kenton, D., Knecht, L., Krasnov, S., Sandomirskiy, V., Sirotnin, V., Starchenko, G., Wilbur, J. & Zipsper, J. (1999). MEDLINE, Vol. 1999. National Library of Medicine (NLM).
58. Bilu, Y. & Linial, M. (2002). The advantage of functional prediction based on clustering of yeast genes and its correlation with non-sequence based classifications. *Journal of Computational Biology* 9, 193-210.
59. (1997). The yeast genome directory. *Nature* 387, 5.
60. Przybylski, D. & Rost, B. (2002). Alignments grow, secondary structure prediction improves. *Proteins: Structure, Function, and Genetics* 46, 195-205.
61. Nair, R., Cokol, M. & Rost, B. (2000). PredictNLS: prediction of nuclear localisation signals, Vol. 2000. CUBIC, Columbia University, Dept. Biochemistry & Mol. Biophysics.
62. Bork, P. & Gibson, T. J. (1996). Applying motif and profile searches. *Methods in Enzymology* 266, 162-184.
63. Rost, B. & Valencia, A. (1996). Pitfalls of protein sequence analysis. *Current Opinion in Biotechnology* 7, 457-461.
64. Eisenhaber, F. & Bork, P. (1998). Wanted: subcellular localization of proteins based on sequence. *Trends in Cell Biology* 8, 169-170.
65. Devos, D. & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in Genetics* 17, 429-431.
66. Pawlowski, K., Jaroszewski, L., Rychlewski, L. & Godzik, A. (2000). Sensitive sequence comparison as protein function predictor. *Pac Symp Biocomput* 8, 42-53.
67. Jones, S. & Sgouros, J. (2001). The cohesin complex: sequence homologies, interaction networks and shared motifs. *Genome Biology* 2, RESEARCH0009.1-0009.12.
68. Felsenstein, J. (1988). PHYLIP: phylogeny inference package. *Cladistics* 5, 355-356.
69. Rigoutsos, I., Floratos, A., Ouzounis, C., Gao, Y. & Parida, L. (1999). Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. *Proteins: Structure, Function, and Genetics* 37, 264-277.
70. Kel, A. E., Kel-Margoulis, O. V., Farnham, P. J., Bartley, S. M., Wingender, E. & Zhang, M. Q. (2001). Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J Mol Biol* 309, 99-120.