Proteome of Caulobacter crescentus cell cycle publicly accessible on SWICZ server

Here we present the Swiss-Czech Proteomics Server (SWICZ), which hosts the proteomic database summarizing information about the cell cycle of the aquatic bacterium Caulobacter crescentus. The database provides a searchable tool for easy access of global protein synthesis and protein stability data as examined during the C. crescentus cell cycle. Protein synthesis data collected from five different cell cycle stages were determined for each protein spot as a relative value of the total amount of [35S]methionine incorporation. Protein stability of pulse-labeled extracts were measured during a chase period equivalent to one cell cycle unit. Quantitative information for individual proteins together with descriptive data such as protein identities, apparent molecular masses and isoelectric points, were combined with information on protein function, genomic context, and the cell cycle stage, and were then assembled in a relational database with a world wide web interface (http://proteom.biomed.cas.cz), which allows the database records to be searched and displays the recovered information. A total of 1250 protein spots were reproducibly detected on two-dimensional gel electropherograms, 295 of which were identified by mass spectroscopy. The database is accessible either through clickable two-dimensional gel electrophoretic maps or by means of a set of dedicated search engines. Basic characterization of the experimental procedures, data processing, and a comprehensive description of the web site are presented. In its current state, the SWICZ proteome database provides a platform for the incorporation of new data emerging from extended functional studies on the C. crescentus proteome.

Keywords: Bioinformatics / Caulobacter crescentus / Cell cycle / Global protein analysis

1 Introduction

Complex regulatory networks operate on the level of differential protein synthesis, modification and rapid protein turnover. Various high-throughput expression techniques such as DNA-microarrays and 2-D protein gels in combination with mass spectrometry (MS) have been developed to exploit the large body of now available sequence data and provide surveys of expression data for thousands of genes or proteins in a single assay. Transcription is only the first step in gene regulation and mRNA is not the ultimate product of a gene. To fully appreciate all possible levels of gene expression control including post-transcriptional and post-translational regulation, careful analysis of the changes occurring at the proteome level are needed. We have recently performed an analysis of protein synthesis and degradation during the cell cycle of the aquatic bacterium Caulobacter crescentus [1]. 15% of the protein spots, which were reproducibly detected on 2-D gels, were subject to differential synthesis during the cell cycle and about one out of twenty proteins was rapidly degraded in the course of one cell cycle [1]. Most importantly, more than 50% of the rapidly degraded proteins were also subject to cell cycle-dependent synthesis, postulating a strong link between cell cycle control and protein degradation in this organism. Here we document a C. crescentus proteome database, which is publicly accessible and equipped with search engines and hyperlinks that connect the data with other existing public databases. The database not only provides protein identification, synthesis and stability data but also allows access to raw data and protocols used for data acquisition.

All data created by Grünenfelder et al. [1], as well as additional protein identifications, were assembled into a relational database and stored in the newly created SWICZ database.
(Swiss-Czech) server accessible through http://proteom.biomed.cas.cz. To enable the communication of users with the database via a www server and a www interface a set of client/server scripts was created. The newly established server currently hosts proteomic data of two organisms, C. crescentus and Streptomyces coelicolor. In this paper we describe the database, the user interface and the features of the C. crescentus part of the server.

2 Materials and methods

Sample preparation, electrophoresis conditions, methods used for MS analysis, and other experimental details were published previously [1]. Here we list only items necessary for the understanding of the database organization and the web interface, which will allow the user to instantly use the web site.

2.1 Cell cycle experiment and protein pulse-labeling

C. crescentus cultures were synchronized by density gradient centrifugation [2], which separates higher density swarmer cells from lower density stalked and predivisional cells. Swarmer cells were released into fresh glucose minimal medium (M2G) [3], and analyzed as they progressed through the cell cycle. Protein synthesis was measured during the cell cycle by pulse labeling cell aliquots of a synchronized culture with \([^{35}\text{S}]\)methionine in the G1 phase (0 cell cycle units), early S phase (0.3), late S phase (0.6), G2 phase (0.8), and immediately following cell division (1.0). To assess protein stability protein samples were analyzed 0, 60 and 120 min after the addition of a chase solution (0.2% tryptone, 1 mM methionine, 0.02% cysteine). Labeled cells were harvested and proteins were resolved after cell lysis on analytical 2-D gels as described [1].

To assess the influence of the synchronization procedure on protein synthesis, expression patterns were compared in asynchronous cultures before and after the synchronization procedure. Swarmer, stalked and predivisional cells were re-pooled after separation by density gradient centrifugation and pulse-labeled as described above. The protein expression pattern was then compared to a labeled culture that had not experienced the synchronization procedure. The marker proteins were visualized by silver staining following the standard protocol [4].

2.2 2-D electrophoresis

In the first dimension, proteins were separated in 18 cm long IPG gels with a non-linear pH gradient ranging from pH 3 to 10 (Immobiline DryStrip pH 3–19 NL 18 cm, Amersham Biosciences, Uppsala, Sweden) in a horizontal electrophoresis tray (Multiphor, Amersham Biosciences). In the second dimension, proteins were separated with standard continuous 12% SDS gels of 18 cm × 18 cm × 1 mm size in a PROTEAN II xi multiecell (Bio-Rad Laboratories, Hercules, CA, USA) using acrylamide with 0.8% piperazylacylamide crosslinker (Acrylogel-PIP, BDH, Poole, UK). The radioactively labeled proteins were detected by exposing the gels to storage phosphor screens (Molecular Dynamics, Sunnyvale, CA, USA) for 48 h. Preparative gels were stained with colloidal Coomassie Blue (Novex, San Diego, CA, USA). Protein size (10–100 kDa) and isoelectric point range (pH 3–10) of the 2-D gels were determined by comigrating 2-D gel marker proteins (Bio-Rad) with a C. protein sample. The marker proteins were visualized by silver staining following the standard protocol [4].

2.3 Data processing

2.3.1 Matchset

Samples from six independent labeling experiments were resolved on six 2-D gels for each time point investigated, forming one experimental set. Scanned gel images were processed with the PDQuest (Bio-Rad) gel analysis software, and the spot intensities were determined for all spots in all gels of the experiment as a fraction of the total amount of labeled proteins. The analyzed gels were assembled to a “matchset” and an artificial reference gel, containing all protein spots which were significantly expressed in at least one experimental set, was created. Corresponding spots in the reference gel and in the member gels of the set were matched using built-in procedures of PDQuest. The matches were checked carefully and corrected where necessary. Data were exported and further processing was done in the S-PLUS statistical software package (MathSoft, Cambridge, MA, USA).

2.3.2 Data clean up

In order to obtain statistically relevant sets of data, the following data processing steps were applied to the cell cycle expression data. 1. Density values were converted...
to ppm by dividing each spot density by total DPM in the gel and multiplied by 10^6. 2. All spots with less than three members in the repeats were removed. 3. Averages of the repeat values were calculated. 4. Normalization for systematic differences [5] was carried out. 5. Outliers were identified using a t-significance test (p = 0.05) and removed. 6. Averages were recomputed and the profiles with spot density data removed by these procedures were deleted from the set.

This procedure selected a minimal set of 1250 protein spots, which were used for further analysis and which were stored in the database.

2.3.3 Cluster analysis
Cluster analysis identified groups of proteins coordinately regulated during the cell cycle. Spot records which exhibited significant change during the cell cycle [defined by analysis of variance (ANOVA), p = 0.01] were selected. In order to remove all proteins whose expression levels were influenced by the synchronization process and other sample treatments, the following set of constraints was applied to the data set – maximum density threshold, synchrony effect, statistical confidence defined by average CV, and a quality threshold comprising reliability of the spot image analysis as defined in PDQuest. All spot profiles with a maximal ppm value in the profile lower than the median of all spot densities were removed. Spot records influenced by the synchronization procedure detected at 0 min and/or at 45 min but not at 140 min, were removed. Spots with an average CV in the cell cycle profile higher than 25% were removed. Spots with an average quality coefficient lower than 20 in the profile were also removed.

The final set of cell cycle-regulated proteins contained 145 spot records. Spots with differential expression patterns were sorted by hierarchical clustering with agglomerative nesting algorithm [6] (Fig. 1), with the distance matrix calculated as 1 - cor(a, b) (Pearson correlation coefficient) from the spot profiles, and from the differences between consecutive members of the spot profile. This combined distance metrics allows spot profiles to be grouped both according to mutual correlation and changes in the protein amounts between different phases of the cell cycle.

2.3.4 MS identifications
Spots were cut out manually from preparative gels, digested with trypsin and the oligopeptide fragments were characterized by MALDI-TOF MS (for details see reference 1 or the experimental procedures section of the web site). Proteins were identified by comparison of the resulting peptides with the whole genome sequence of C. crescentus [7, 8].

3 Results and discussion
3.1 Database and WWW interface
Proteomic studies generate large amounts of data that need to be stored in easily accessible, interlinked databases to be useful for the scientific community. Here we describe a 2-DE based proteome database for the microorganism C. crescentus. The database stores previously published data on global cell cycle-regulated protein synthesis and stability [1] as well as the identity of 295 2-D gel spots, many of which have not been previously reported.

The database can be accessed via the SWICZ server (http://proteom.biomed.cas.cz), which also hosts proteome data for the microorganism S. coelicolor. Access to the Caulobacter main menu is through the “Caulobacter crescentus” button on the home page of the SWICZ server.

From the main menu the user can gain access to gel images, query pages, and to the public area, where experimental protocols, news and messages concerning the database, the user interface and the server can be obtained. The directions of hyperlinks among individual pages are shown in Fig. 2. From any page, the user can return to either the main server or the Caulobacter home pages. A search for a protein with a known position on the 2-D gel can be initiated on clickable 2-D gel maps. Alternatively, other criteria like spot number, molecular weight, pI range, protein function, or cell cycle expression cluster can be used to submit database queries.

3.2 Access to the basic record
Two main entry points for the access to the records of the database exist:Clickable 2-D gel maps (Fig. 3) and database query forms, listed in the windows “Clickable Maps” and “Searches”, respectively. Two different 2-D gel maps provide links to the records: The first map shows an example of an analytical gel and represents an autoradiogram of labeled proteins corresponding to 10^6 DPM of [35S]methionine and cysteine. The second map shows an example of a preparative gel (1 mg total protein) stained with Coomassie brilliant blue. This type of gel was used for the isolation and identification of protein spots by MS.

Database records can be accessed simply by clicking on the desired spot on one of the maps. A server script creates a database query and the corresponding database
Figure 1. Hierarchical clustering tree of cell cycle-dependent spot expression profiles. Hierarchical clustering divides the set into seven main and 23 individual clusters. Members of clusters can also be found on the SWICZ web site by means of a search engine. Details about clustering of proteins with a cell cycle-dependent synthesis profile can be found in the supplementary materials of Grünenfelder et al. [1].
Figure 2. Schematic description of the web site layout of the C. crescentus proteome server. Arrows indicate connections between different pages. From each page the user can directly return to the Caulobacter home page. The “Public area” which contains protocols, and information about the server is not shown in this figure.
Figure 3. Analytical (left, based on autoradiogram of $10^6$ DPM of $[^{35}S]$methionine and cysteine labeled proteins) and preparative (right, based on Coomassie stain of 1 mg total protein) 2-D electrophoretic gels. Both gels can serve to access the individual spot records corresponding to the identified proteins. By clicking on any of the marked spots, the corresponding spot record is invoked.

The position of each identified protein on the reference gels can be displayed by clicking on the two hyperlinks. The bar graphs indicated below the query output show changes in the rate of protein synthesis during the cell cycle, and changes in protein levels during the chase period as determined by pulse-chase experiments, respectively.

The output for each 2-D gel spot contains 19 records and two bar graphs. The first item is a unique spot identifier. Items 2–5 contain the coordinates of the protein spots on the analytical and preparative gels, respectively. Items 6 and 7 give basic physico-chemical characteristics of the respective proteins, i.e., calculated molecular mass and $pI$. Item 8 gives the ORF reference number, item 9 contains a direct hyperlink to the respective record in GenBank. By clicking on the hyperlink a database query is sent to http://www.ncbi.nlm.nih.gov/ and a new window opens the corresponding GenBank record. Items 10 and 11 provide information about the protein function. Item 12 describes the method which was used to identify the respective protein spots. In item 13 protein spots with a differential cell cycle synthesis pattern are marked. Spots oscillating during the cell cycle are indicated as “differential”. No entry indicates that the synthesis of the respective protein is not significantly influenced by the cell cycle (for the details about selection of cell cycle-controlled proteins see Section 2.1). Item 14 lists the cell cycle expression cluster number (Fig. 1) and gives a hyperlink to other members of the same cluster. Clicking on the hyperlink invokes a list of members of the cluster with links to their database records. Item 15 compares the cell cycle-dependent protein expression patterns with the corresponding mRNA transcription data from a parallel DNA microarray study by Laub et al. [9]. “Yes”: Cell cycle-dependent fluctuation of mRNA levels and protein synthesis match. “No”: Protein synthesis but mRNA levels do not fluctuate during the cell cycle. These are examples of proteins, which are subject to cell cycle control on a post-transcriptional level. “Inverse”: mRNA levels and protein synthesis profiles behave in an inverse manner. “No data”: No data are available from the transcriptome study. Item 16 lists examples where the corresponding mRNA data was assigned to a different ORF number in the transcriptome analysis. Under item 17, proteins whose synthesis was affected more than two-fold by the synchrony procedure, are listed. An increase in protein synthesis by the synchronization procedure is indicated as “induction”; A decrease in protein synthesis...
Figure 4. Example of a typical spot record. A list of features associated with a particular spot is indicated. The hyperlink “GenBank reference number” (item 9) invokes the GenBank record of the identified protein. The hyperlink “Cell cycle expression cluster” (item 14) invokes a list of members of the same cell cycle expression cluster (Fig. 1, if the field is empty, the protein is not differentially regulated during cell cycle). “Cell cycle-dependent transcription” (item 15) cross-references genes which were identified in a parallel study as being subject to cell cycle-dependent transcription.

as a consequence of synchronization is marked by “repression”. Item 18 refers to the protein stability by listing the change in protein abundance during a pulse-chase experiment as “increase”, or “decrease”. An empty field refers to changes in protein stability, which were not statistically significant. Item 19 contains a reference to N-terminal export signal sequences predicted by the SignalP algorithm (http://www.cbs.dtu.dk/services/SignalP/ [10]).

The two bar graphs below the list of output records depict expression levels as a function of the cell cycle (first graph), or protein levels during the chase period (second graph). The time in min refers to the progression of the
3.3 Search engines

The web interface allows searching for a spot record or a set of records in six basic and seven additional database query items accessible on the Caulobacter main page in the window “Searches”. These are spot number, range of values of molecular weight and pI, protein functional group of the protein, description field, cell cycle expression cluster as defined by cluster analysis, and ORF number. Additional searchable criteria are grouped under the “other criteria” item, and include the method of identification, cell cycle expression, cell cycle-dependent transcription, synchrony effect, protein stability (change during pulse-chase), export signal peptide, and GenBank identification number. In each case either a spot record page is displayed or a list of spots with hyperlinks to the individual spot records is brought up.

3.4 Cache

Each single spot record can be temporarily stored in a cache by clicking on the “Add to cache” button on top of the record page. The list of records can be viewed and edited in a separate page. The records are exported in an MS Excel compatible format. The page can then be copied directly to an MS Excel worksheet.

3.5 Linking to SWICZ

A direct link from a remote server to selected search engines can be made. The rules can be found on the Caulobacter home page under the item “Linking the database” in the window “Public area”, or directly under http://proteom.biomed.cas.cz/caulob/cclink.php. The general syntax is http://proteom.biomed.cas.cz/queryswicz.php?dbase=XXXX&query=XXXX&n1=XXXX&n2=XXXX&n3=XXXX&n4=XXXX. The explanation of the individual parameters of the query can be found on the page and they differ for the two databases (Caulobacter and Streptomyces) hosted by the server.

3.6 Web server functionality

The SWICZ is hosted by a stand-alone server powered by Microsoft Windows NT 4.0 Server. Data are stored in a Microsoft Access 2000 database and accessed by means ODBC and standard SQL queries. Selective database searches are mostly based on the SQL clause SELECT <select list> WHERE <search condition>. The server-side scripts are written in PHP 4.0 (see the online manual http://www.php.net/manual/en/ for detailed description). The scripts may be divided into several groups: those accessing data most frequently use PHP functions ODBC_Open(), ODBC_Exec(), ODBC_Result(), and ODBC_Fetch_Row(); scripts processing user’s input employ functions SubStr(), StrVal(), StrLen(), StrStr(), Round() and the like; scripts that work with graphic information (e.g., mark certain protein spots on electrophoretograms with crosses, print protein numbers etc.) are based on functions Header(“Content-type: image/gif”), ImageCreateFromGIF(), ImageColorAllocate(), ImageLine(), ImageString(), ImageGIF(), and so on. Protein cache is supported by session functions Session_Start() and Session_Register(). Client-side scripting used e.g. in form validators is written in JavaScript.

3.7 Future perspectives

New data on protein identity, abundance, stability, modification, etc. will be continuously incorporated into the existing Caulobacter protein database. Assignment of protein functions will be made according to the COG database (http://www.ncbi.nlm.nih.gov/COG/) standard. To guarantee maximal flexibility and transparency of the available expression and stability data, the proteome database will eventually be linked directly to the Caulobacter transcriptome database (http://caulobacter.stanford.edu/CellCycle/). Combining the search engines will allow the user to create more complex Boolean searches. Authorized users will also have a direct access to the SQL query editor.

4 Concluding remarks

The SWICZ web server hosts a unique proteomic database of C. crescentus, which contains data on global changes of protein synthesis and stability during the cell cycle as well as the identity of currently 250 protein spots localized on 2-D gels. The web design allows a user-friendly examination of the database and the analysis of the results by means of a set of search engines, outputs, and a cache. Links to other data sources such as GenBank were made wherever possible. Complementary data such as protocols, manuscripts concerning the database, and raw data are stored in the “Public area” of the Caulobacter home page from where they can be accessed and downloaded.
The work was supported by Swiss National Science Foundation fellowship 31-59050.99 to U. J., by a grant from Hoffmann-LaRoche, Ltd., to U. J. and by Institutional Research concept no. AVOL 502090, grants 310/03/0293 of GACR and IAA5020211 GACAS to J.V.

Received December 6, 2002

5 References