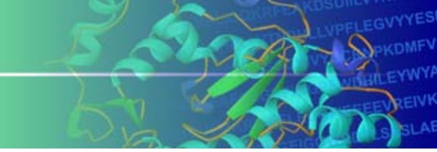

Center for Eukaryotic Structural Genomics

Protein Structure Initiative

Current cost analysis of structure determination at the Center for Eukaryotic Structural Genomics

John G. Primm, Craig S. Newman, Eldon L. Ulrich, Zhaohui Sun, Russell L. Wrobel, Ronnie O. Frederick,
Hassan K. Sreenath, Frank C. Vojtik, Min S. Lee, Craig A. Bingman, Dmitriy A. Vinarov, Zsolt Zolnai,
David J. Hruby, Donna M. Troestler, Brian F. Volkman, Brian G. Fox, George N. Phillips Jr., and John L. Markley

University of Wisconsin-Madison, Department of Biochemistry, 433 Babcock Drive, Madison, Wisconsin, USA 53706-1549, http://www.uwstructuralgenomics.org

Abstract

Cost predictions, analysis, and control are vital aspects of the Protein Structure Initiative (PSI), Phase 1, project at the Center for Eukaryotic Structural Genomics (CESG). In response to inadequacies in University of Wisconsin-Madison (UW-Madison) accounting practices and the need to analyze efficiencies within our protein production pipeline, we have developed a basic real-time accounting system that tracks costs by teams within our project. By combining this information with information gathered from Sesame, our laboratory information management system (LIMS), we are able to chart the output of each team over time and now have sufficient hard financial data to evaluate the costs associated with each step of protein production and structure determination. Costs which are tracked include materials, labor, services, administration, general overhead, and production activities. Costs charged to research and development activities are also considered. Here we present a breakout of these factors by team and a build-up of the costs associated with structure determination at CESG.

Introduction

Official financial information for CESG's is now based on the UW-Madison's *Shared Financial System (SFS),* which is an implementation of *PeopleSoft Financial Software for Higher Education*. Information is delivered to users via WISDM (WISconsin Data Mart), a web-based financial reporting system (**Figure 1**). The typical scenario for UW-Madison departments funded by grant money is that all costs associated with the work, whether labor or purchases, will be funded directly on a particular account established to support the grant; essentially, one account per grant. The system is also based on batch processing. While the system has strengths at the university level, it presents accounting challenges at the individual P50 grant level.

Figure 1. The University of Wisconsin-Madison's WISDM system allows users to perform various project/grant-related data searches and reports, including search projects by PIs, by department, by fund, or by a range of project/grant IDs. It enables users to create salary encumbrance balance reports and purchase order balance reports for projects/grants.

WISDM presents significant hurdles to managing the funds of large dynamic grants, and, specifically, cannot handle a variety of requirements encountered with PSI-1 grant funding. CESG has an average business day expense in excess of $27,000, but the WISDM system, coupled with the university bureaucracy, creates significant latencies in the billing information being available in the system. It also does not allow for accounting by sub-tasks within CESG's production pipeline or for easily maintained division of sub-budgets and modeling projected expenses. It does not allow for real-time accounting and real-time performance management of structure determination by the pipeline.

PSI-1 funding requires increased accountability for activities, and reporting on many sub-tasks within the overall program may be requested. The effort to reduce the cost per structure by NIGMS/NIH requires real-time accounting and real-time performance management. Real-time performance management demands: 1) focus on optimum potential, 2) operationally-focused cost and profitability measures for resource allocation decisions, 3) dynamic performance targets that reflect changing definitions of what is important and what is acceptable performance, 4) real-time monitoring of performance, 5) external factors to keep the organization focused on the "right things," and, 6) use of continuous improvement as the overall driver for organizational development.

CESG Pipeline

The CESG pipeline is currently divided into two separate primary branches (**Figure 2**). During initial selection, targets are sorted into groups which are sorted by size. The smaller targets enter the Cell-Free Production pathway, and the larger targets **are** routed to *E. coli*-based Cell Production. As a secondary pathway, we have created salvage workgroups comprised of targets that have failed in the *E. coli* branch which are then attempted in the cell-free branch, and visa versa. As cell-free and *E. coli*-based production can confer different solubility expression characteristics on a given target, this offers an attractive alternative to increase overall yield per workgroup of selected targets [1].

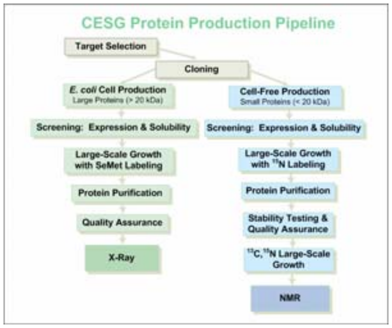
CESG Protein Production Pipeline
Target Selection
Cloning
E. coli Cell Production
Large Proteins (> 20 kDa)
Cell-Free Production
Small Proteins (< 20 kDa)
Screening: Expression & Solubility
Screening: Expression & Solubility
Large-Scale Growth with SeMet Labeling
Large-Scale Growth with ^{15}N Labeling
Protein Purification
Protein Purification
Quality Assurance
Stability Testing & Quality Assurance
X-Ray
^{13}C,^{15}N Large-Scale Growth
NMR

Figure 2. The primary branches/pathways which targets follow in the current configuration of CESG Protein Production Pipeline.

Tracking Pipeline Progress

In order to track the progress of CESG's pipeline, we combine data acquired from WISDM and from Sesame **[2].** Sesame tracks a number of actions (**Figure 3**), which are entered by individual team members. These actions correspond to distinct outputs for each team which is responsible for specific activities of the pipeline.

Sesame Definitions of Actions

Highlighted actions should result in a "work stopped" action being added immediately afterwards

Highlighted actions represent a "seal of approval" from that group

Group	Action	NIH Tag	Definition
Bioinformatics	Selected	Selected	ORF has been added to a pipeline work group
	Primers Ordered		Primers ordered from vendor
Cloning	Primers Received		Primers received from vendor
	PCR +		Band of approximately the correct size visible on the gel
	PCR -		No band visible on the gel
	Entry Clone +		Band visible on the PCR screen
	Entry Clone -		No band visible on the PCR screen
	Cloning Acceptable	Cloned	Destination clone ready for expression testing is available
Expression Testing	Screening Expression +		A distinctive or strongly expressing protein band is clearly visible on the SDS-PAGE gel
	Screening Expression -		No distinctive or strongly expressing protein band is seen on the SDS-PAGE gel.
	Screening Acceptable	Expressed	Target protein clone is suitable for large scale protein production, and therefore selected for subsequent transformation and transfer to next section.
	Transformed		Target clone DNA is used in transformation (into one of two E coli strains; B834 or Rosetta), and will be passed onto large scale protein production.

Figure 3. Example of some 'Sesame Actions'. CESG's data dictionary currently defines 195 separate actions. Those shown here are the subset associated with target **selection, cloning and expression testing**. Note that some actions lead to "Work Stopped" status for a particular target through the protocol being followed. "NIH Tags" correspond to information collected by TargetDB.

CESG Real-Time Accounting

In order to track and control the budget and performance of CESG, we have instituted a Microsoft Excel-based real-time accounting system. The Supply and Services funds, designated as "Other," are divided into 12 team-specific sub-accounts in order to better track project costs associated with each step of the CESG protocols on a monthly and quarterly basis (**Figure 2**). The CESG team accounts are updated as orders are routed from CESG through the University apparatus. The system also contains forward projections for each team's expenses. Team leaders are responsible for individual monthly budgets, with tactical spending control in the hands of the CESG Project Manager. The overall goals and research plans for CESG are set by the CESG Executive Committee and Finance Committee. Specific team budgets and projections are reviewed on a regular basis with CESG Team Leaders, the CESG Finance Committee, and the UW-Madison administration.

In addition to cost break-outs by functional team, the CESG system combines the expenses into NIH and institutional accounting categories for purposes of account reconciliation and communication with these entities (**Table 1**). *At this time point, the total project budget for these combined areas was < 2% different than expected.*

Table 1. Supplies and Other (i.e., Services) monthly budget by team for December 2004. Fiscal year-to-date statistics are also tracked. Some significant over and under spent team budgets are seen. These individual budgets will be re-balanced to reflect the project priorities as set by the Executive Committee.

	Total Project	Administration	Bioinformatics	Cell-Free	Cloning	Computing	Exp. Testing	Large E. coli	NMR	Purification	QA	Sesame	X-ray
December-04	144,500.00	36,000.00	1,000.00	22,000.00	30,000.00	2,000.00	4,000.00	6,000.00	25,000.00	10,000.00	2,000.00	1,000.00	5,500.00
Carryover from November 2004	33,459.31	13,881.95	1,023.51	102,230.90	19,518.38	246.23	1,200.17	3,556.21	45,815.38	11,341.63	1,993.17	145.70	5,709.44
Budget to date	111,040.69	22,118.05	2,023.51	80,230.90	49,518.38	1,753.77	5,200.17	2,443.79	70,815.38	21,341.63	3,993.17	854.30	11,209.44
Spent this month:	91280.6	33222.09	0	8971.82	18144.44	3500	490.2	4633.1	12171.29	1891.17	629.63	0	7626.86
Current Balance:	19,760.09	11,104.04	2,023.51	89,202.72	31,373.94	1,746.23	4,709.97	2,189.31	58,644.09	19,450.46	3,363.54	854.30	3,582.58
Total budget	1,251,000.00	213,000.00	10,500.00	218,000.00	232,000.00	19,500.00	40,000.00	64,000.00	252,000.00	108,000.00	22,000.00	12,000.00	60,000.00
Year to date	530,239.91	104,104.04	1,976.49	177,202.72	97,626.06	9,746.23	11,290.03	26,189.31	55,355.91	20,549.54	4,636.46	3,145.70	18,417.42
% spent of total budget available	42.39%	48.88%	18.82%	81.29%	42.08%	49.98%	28.23%	40.92%	21.97%	19.03%	21.07%	26.21%	30.70%
Target % spent	43.96%	43.66%	38.10%	40.37%	55.60%	41.03%	40.00%	37.50%	45.24%	37.04%	36.36%	33.33%	36.67%
% spent - Target %	1.58%	-5.21%	19.27%	-40.92%	13.52%	-8.96%	11.77%	-3.42%	23.27%	18.01%	15.29%	7.12%	5.97%

To illustrate the costs associated with operation of, and investment in, the CESG pipeline, we show the four-year budget with the traditional budget categories (**Figure 4**). During the first years of the project, significant investment was made in equipment, while during the later years personnel, supplies, and indirect costs were the dominant factors in the budget.

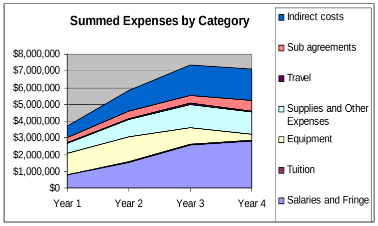

Figure 4. CESG budget expenses by traditional budget categories through the four fiscal years of the project.

A further breakdown of the budget by Team per calendar quarter is possible using this accounting system (**Figure 5**). While the real-time system captures costs on a daily basis, and reports are available by month, a longer time frame is needed to accumulate meaningful data due to the nature of supplies ordering and pipeline material handling. In this model the sub-agreements are distributed to the teams most appropriate. We currently have four sub-awards with defined deliverables to the NMR and Bioinformatics Teams.

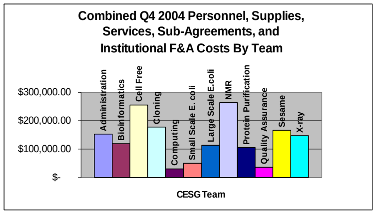
Combined Q4 2004 Personnel, Supplies, Services, Sub-Agreements, and Institutional F&A Costs By Team
$300,000.00
$200,000.00
$100,000.00
$-
Administration
Bioinformatics
Cell Free
Cloning
Computing
Small Scale E. coli
Large Scale E.coli
NMR
Protein Purification
Quality Assurance
Sesame
X-ray
CESG Team

Figure 5. Actual charges by team during the fourth quarter of the calendar year 2004.

Data Mining and Reporting

The Sesame LIMS system is designed primarily as a tool to capture pipeline activities. However, provisions were made in the design of the LIMS to capture non-pipeline research activities to improve the pipeline as well. The limits of capturing research and development activities include the structural problem that the new activity may truly be unique, and therefore not described as a specific action in the system. Yet, at the same time, many development activities must include many routine steps in order to accomplish the unique development goal. Therefore distinctions between pipeline and non-pipeline activities, as captured by the LIMS, can qualitatively capture the projects relative effort development versus pipeline over time (**Figure 6**).

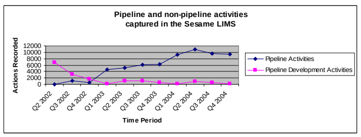
Pipeline and non-pipeline activities captured in the Sesame LIMS
Actions Recorded
12000
10000
8000
6000
4000
2000
0
Q2 2002
Q3 2002
Q4 2002
Q1 2003
Q2 2003
Q3 2003
Q4 2003
Q1 2004
Q2 2004
Q3 2004
Q4 2004
Time Period
Pipeline Activities
Pipeline Development Activities

Figure 6. Pipeline development activities (research and development) versus pipeline activities over the course of PSI-1. As we move forward, we anticipate another burst of development activity during the remainder of PSI-1.

Significant cost, time, and energy has been devoted to creating a truly functional, scalable pipeline. Research effort to develop the PSI-1 pipeline has now come to fruition with most resources being devoted to production during the second half of calendar year 2004. We consider that during the last three months of calendar year 2004 our phase-one pipeline configuration operated **near** peak efficiency. As we go forward, research activities will again increase for a period of time as we re-tool our pipeline for micro-scale screening. (Optimization experiments are in progress.)

We mine actions logged into Sesame and produce time-based reports of cell-based pipeline activity and cell free pipeline activity. It is important to note that all reported activities are sorted into two types: "work-based" and "target-based" actions. The "work-based" reports generated reflect total effort put into the pipeline. This is an important measure of effort versus cost. Perhaps more important is the ability of the software to generate "target-based" reports which take into account the multiple threads a particular target may take through the pipeline, but will only count a single positive event for a unique target at any given stage along the CESG pipeline (**Table 2**).

Table 2. **Cumulative** target-based report for CESG as a whole.

Stages	Q2 2002	Q3 2002	Q4 2002	Q1 2003	Q2 2003	Q3 2003	Q4 2003	Q1 2004	Q2 2004	Q3 2004	Q4 2004
Selected Targets	904	1718	1908	2007	2478	3006	3144	3999	4752	5464	5887
PCR +	362	1053	1338	1432	1716	2008	2269	2631	3480	4022	4638
Entry Clone +	167	646	937	1086	1318	1556	1800	2116	2893	3420	3940
Sequence +	167	550	703	1079	1173	1421	1646	2023	2678	3221	3752
Destination Clone +	140	407	545	892	948	1239	1462	1821	2400	2781	3355
Screening Expression +	80	125	211	472	535	699	842	1067	1386	1743	2018
Screening Soluble +	26	65	121	254	391	502	701	905	1178	1519	1780
Large-Scale Cell Growth +	1	24	73	190	303	413	613	823	1097	1438	1699
Production Scale Expression +	1	12	20	81	225	339	535	735	1000	1342	1603
Soluble Production +	1	12	20	55	166	271	440	619	849	1112	1333
Tag Cleaved +	1	12	18	26	105	190	262	324	567	729	794
Purification Completed	1	2	6	13	40	109	163	188	391	506	567
Successfully Purified	1	2	5	9	21	58	107	144	239	343	384
To Crystallization Screening	1	2	5	5	8	28	85	109	178	243	266
Crystallized +	1	2	4	4	7	9	36	57	68	76	90
Diffraction Quality Crystal +	0	1	2	2	4	6	9	14	22	28	36
Xray PDB Deposited	0	0	0	0	0	1	4	4	10	17	28
HSQC +	0	1	4	6	6	9	9	16	21	44	52
NMR Structure	0	0	0	2	2	3	3	5	7	11	18
NMR PDB Deposited	0	0	0	2	2	2	3	3	7	7	16
BMRB Deposited	0	0	0	1	1	1	2	2	7	7	16

Cost Build-Up of Structure Determination

By using data captured from the CESG real-time accounting system and the time based reports generated by the bioinformatics data mining and reporting, we are able to analyze our output by team and on a project wide basis. The project shares some resources, and we find it helpful break these into a separate category to be sure that the costs are recognized. These resources are the Sesame (LIMS), bioinformatics, computing core, and the administrative teams. The other major categories that are combined from the other reports are the personnel costs for the hands-on pipeline teams (these include fringe benefit costs), the supply and services budget, and the institutional overhead costs.

CESG has invested in a scaleable pipeline using current technology and directed the resources of the project to running this pipeline configuration as evidenced by the increase in pipeline activities during calendar 2004. Our goal for year four was to decrease the average cost per structure to ~$120,000. During the last quarter of the year, a significant decrease in the average cost per structure was seen as the cost decreased from approximately $133,000 to $89,000.

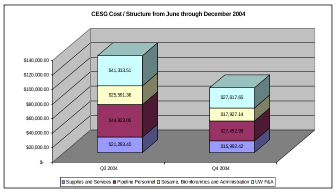

Figure 7. This shows the average cost per structure during the third and fourth quarters of 2004.

Efforts at CESG are currently underway to significantly decrease the cost per structure by streamlining NMR data collection and structure determination, using microscale x-ray crystallography screening techniques, to unifying our cell-based and cell-free cloning strategy, and by increasing our ability to predict large-scale *E. coli* growth success, to improve small-scale expression and solubility screening stages.

References

1. Vinarov, D.A., Lytle, B.L., Peterson, F.C., Tyler, E.M., Volkman, B.F., and Markley, J.L. (2004) Cell-free protein production and labeling protocol for NMR-based structural Proteomics, *Nature Methods*, **1**, 149-153.

2. Zolnai, Z., Lee, P.T., Li, J., Chapman, M.R., Newman, C.S., Phillips, G.N., Jr., Rayment, I., Ulrich, E.L., Volkman, B.F., and Markley, J.L. (2003) Project management system for structural and functional proteomics: Sesame, *JSFG*, **4**, 11-23.

CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative (P50 GM 64598)