Hubert Lam 12-13-01
MB&B 752a Final Project

Understanding Polar Differentiation through Bioinformatics

In the world of bacteria, the two component system of signal transduction has been discovered as a fast and efficient way of sensing external stimuli. Recently, the two component system has also been implicated in a number of processes irrespective of the external environment such as the control of the cell cycle. Such is the case with the asymmetrically dividing bacteria, *Caulobacter crescentus*, where a crucial two-component like pathway has been shown to be involved in polar differentiation and cell division (Hecht 1995). This pathway involves two histidine kinases, PleC and DivJ, having a role in the regulation of an essential response regulator DivK. An interesting discovery in this pathway has been that these cell cycle proteins show a specific and faithful spatial and temporal localization during the cell cycle (Jacobs 2001). Until now the mechanism of localization by these proteins, as well as the function of protein localization has been largely unknown. We have now discovered that localization of the essential DivK response regulator is controlled by phosphorylation at a conserved aspartate residue, and this localization is essential for normal differentiation and cell division in *Caulobacter*.

Protein localization being controlled by phosphorylation in *Caulobacter* is an exciting discovery, but is of little concern for those working outside of the *Caulobacter* field. Therefore we would like to address if this is a global mechanism of controlling localization in bacteria. To this end, it would be beneficial to look in a number of bacteria with homologs to DivK to see if 1) the function of DivK is conserved and 2) to see if localization and the mechanism of localization are conserved. Bioinformatics is a rapidly rising field in which "biology is conceptualized in terms of macromolecules and then "informatics" techniques (derived from applied maths, computer science, and statistics) are used to understand and organize the information associated with these molecules on a large scale (Luscombe 2001). I will discuss the use of bioinformatics tools, such as Blast, Multiple Sequence Alignment, and Secondary Structure prediction to increase our knowledge of the essential response regulator protein DivK and its associate homologs.

Blast is a similar program to FastA in that it is designed for speed, but is still able to provide specificity. These programs break up a query sequence into a hash table of short words, which allows the hashes to be quickly screened through a given database to find "quick hits" of significant matches. Blast minimizes the time spent on sequences with little similarity by searching for segment pairs which have a score of at least some given value T (Altschul 1990). Any such hit is extended to see if it is contained within a high-scoring alignment. This extension takes up most of the time of Blast. The first Blast program created did not allow for gaps in the match sequence. However, Blast2 or Gapped Blast, which is the Blast program that NCBI (NCBI-BLAST 2001) currently uses, does allow for gaps, but does not decrease the speed of the search (Altschul 1997). It achieves this by requiring two hits (the existence of two non-overlapping word pairs of a close proximity) before extension is allowed. Since the extension takes most of the time, this allows for computational time to decrease. Gapped Blast also implements a new gapped alignment algorithm which uses dynamic programming to extend a central pair of aligned residues in both directions, but dropping those alignments that can't achieve a specified score. These major changes allow for a gapped blast, which does not decrease in performance.

By searching a number of databases which all employ the Blast program (NCBI-BLAST 2001; Scranton-WIT 2001; TIGER 2001), I was able to compile a table with the closest homologues to Caulobacter DivK (Table 1). Almost all of these homologues come from bacteria which are contained in the subdivision of proteobacteria, α-proteobacteria. Only one homolog found from these searches was found to be in another subdivision, the γ-proteobacteria subdivision. This was not surprising because Caulobacter is a member of the α-proteobacteria. Such strong homology (reaching 78.6% protein identity) suggests a conserved mechanism of action and conserved function of these homologs to DivK.

DivK

Species	Identity	Similarity	Reference
Caulobacter crescentus	-	-	TIGER/NCBI
Brucella melitensis	76.0%	86.4%	Scranton-WIT
Brucella suis	76.0%	86.4%	TIGER
Brucella abortus	75.2%	85.6%	Scranton-WIT
Mesorhizobium loti	76.8%	84.8%	TIGER/NCBI
Sinorhizobium meliloti	73.6%	84.0%	NCBI
Rhizobium leguminosarum	73.6%	84.0%	Scranton-WIT
Rhodopseudomonas palustris	69.6%	80.8%	Scranton-WIT
Methylobacterium extorquens	72.2%	82.6%	Scranton-WIT
Agrobacterium tumefaciens	72.8%	83.2%	NCBI
Pseudomonas syringae	49.6%	60.0%	TIGER

Table1: Homologs of Caulobacter DivK. The homologs were found using a Gapped BLAST program and searching the NCBI, TIGER and Scranton-WIT databases

Another bioinformatics tool which is commonly used is Multiple Sequence Alignment. This is a process by which a number of different sequences can be compared at the same time. This process is beneficial because it allows the bioinformaticist to view conserved regions of the protein (or nucleic acids) of interest. These regions are likely to be essential in the function of the protein/gene. The program that was used to construct the multiple alignment was Clustal W(Alignments 2001) and viewed through Boxshade (Boxshade3.21 2001). A multiple alignment is carried out in 3 stages: 1) all the sequences are compared to each other one by one (pairwise alignment) through dynamic programming. However, if dynamic programming was used to construct the multiple alignment all at once it would be extremely time consuming and complex. For this reason, 2) a dendrogram (like a phylogenetic tree) is constructed which pairs the most similar sequences first, describing the approximate groupings of the sequence similarity and 3) the final multiple alignment is carried out using the dendrogram as a guide, so that the most similar get paired first and so on(Thompson 1994).

The multiple alignment of the DivK homologs shows what the Blast searches suggested (Fig. 1). There is a great deal of sequence similarity between these homologues. Fittingly, *Pseudomonas syringae* which is the only bacteria which is not part of the α-proteobacteria is the homologue with the weakest similarity. From this

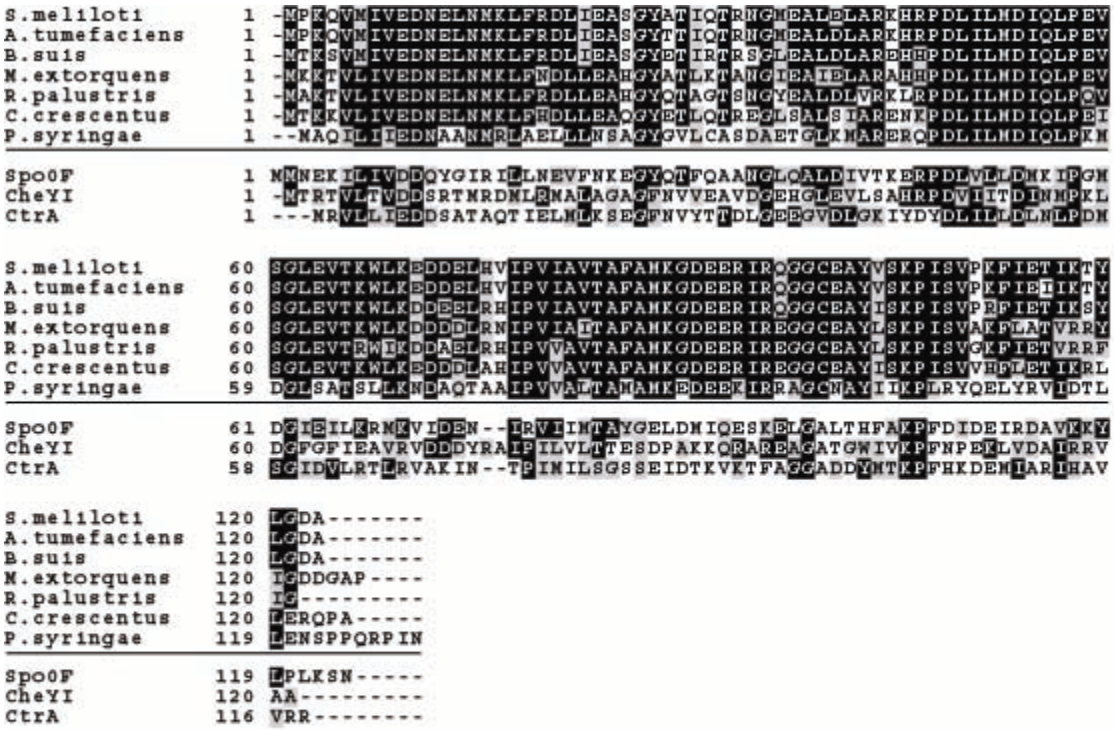

Figure 1: Multiple alignment of DivK homologs and three general response regulators Spo0F, CheYI and CtrA. Viewing alignment of homologs versus the general response regulators shows the similarity is not simply due to nature of response regulators.

multiple alignment, it is evident that there are many areas of the protein with strong homology and it will be interesting to see which of these regions are those that are on the outside of the protein (which would correspond to those that are involved in protein-protein interaction) and those that are on the inside of the protein.

To begin to understand the structure of DivK, secondary structure prediction can be employed. The program PSIPRED was used to predict the secondary structure of all the homologs of DivK and DivK itself (PSIPRED 2001). This prediction method generates a sequence profile (what residues are likely to be found in an alpha-helix, B-sheet, coil-coil) using PSI-Blast (Altschul 1997). It takes advantage of the fact that PSI-Blast naturally creates sequence profiles when it is doing its iterations to get the best sequence comparison. PSIPRED takes the resulting profile and enters it into a simplified neural network (Jones 1999). Neural networks are programs which generalize and learn patterns in a sequence. From the first neural network, a secondary structure is predicted. By running the sequence through a second neural network, this structure can be filtered.

Secondary structure prediction is not an entirely accurate science, however. Because the prediction is based on profiles, it can be largely influenced by database biases, and therefore, a carefully curated database is suggested. Also, as the data for DivK shows (Fig 2.), the evaluation of multiple sequences is complicated because the prediction is a consensus for family members. Thus, the secondary structure for these homologs are implicitly going to be similar and thus interpretation is of differences is difficult.

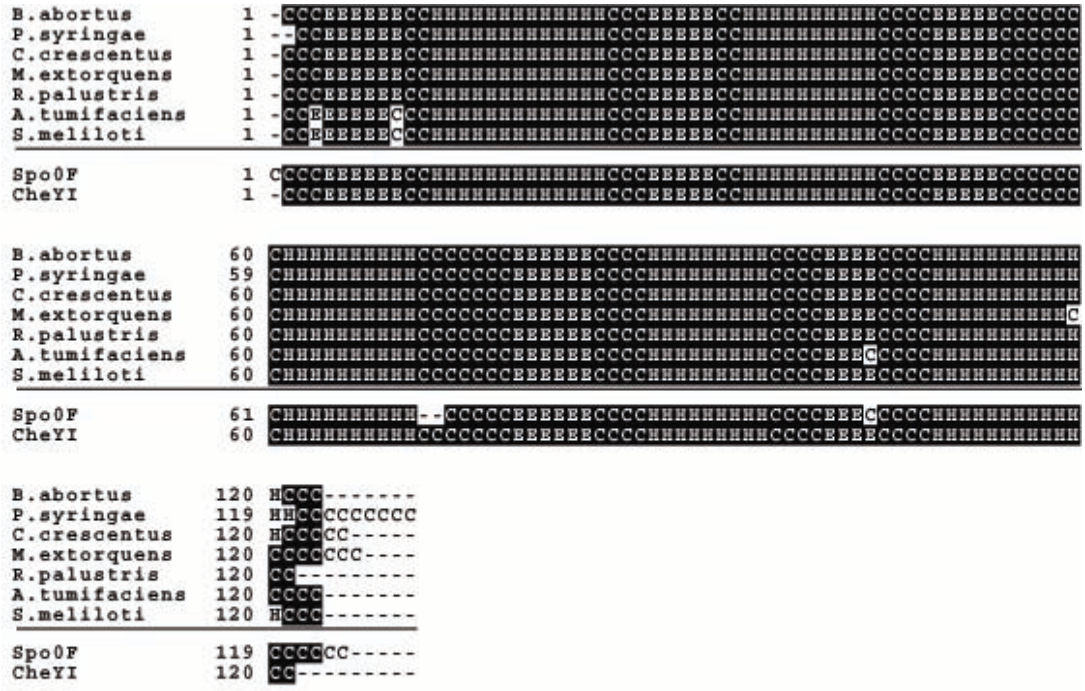

Figure 2: Multiple alignment of Secondary Structure of DivK homologs including two general response regulators: Spo0F and CheYI. Shows that secondary structure prediction for response regulators is very similar, whereas sequence can vary a great deal.

Even the analysis of other response regulators, with weak sequence similarity, have the same secondary structure prediction (Fig 2): Spo0F and CheYI. This similarity, however, can provide a wealth of knowledge about a potential protein-interaction region of DivK. CheY has been shown to interact with a protein FliM in the regulation of chemotaxis (Lee 2001). It does so at the region of α4–β5–α5. By using the comparison of DivK to CheY (Fig 3), we are able to observe that DivK also has a α4–β5–α5 region which could possibly be a protein interaction region for its own partner. This region can be tested by alanine mutagenesis to observe what phenotypes arise.

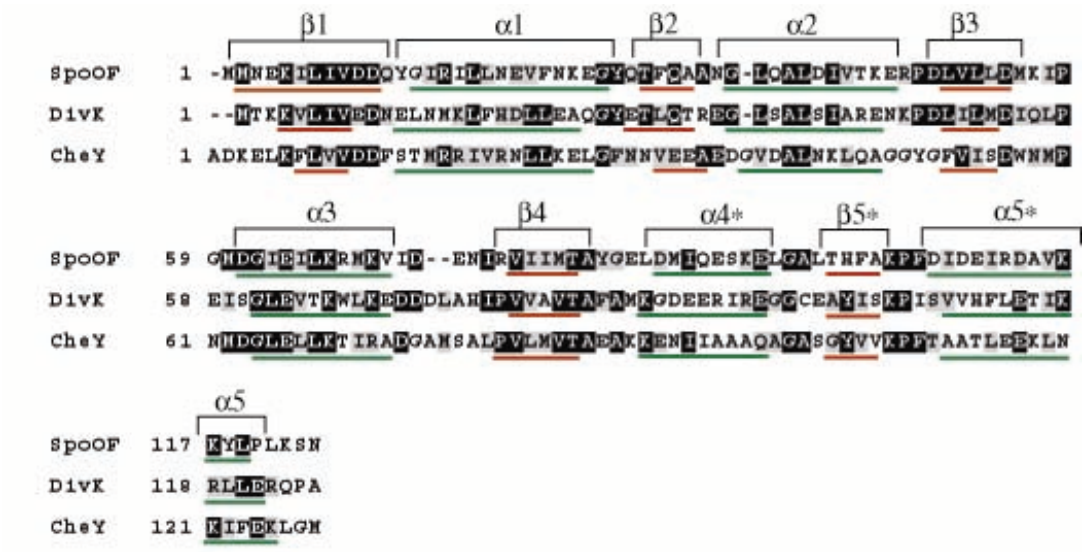

Figure 3: Alignment of Response Regulators: SpoOF, DivK and CheY showing sequence diversity, but secondary structure similarity.

*CheY interaction with FliM occurs at α4–β5–α5

Analysis of tertiary structure, is likely to be the key to discovering the regions essential to function and protein-protein interaction for DivK. A careful analysis of residues which are located on the protein surface, but are unique to the DivK homologs, and not to general response regulators is likely to reveal regions that are specific to DivK protein-protein interactions. The use of bioinformatics to this point has opened up a large area of study apart from *in silico* work. It will be interesting to investigate the close (and distant) homologs of DivK to observe if they localize at the poles of these other bacteria, if localization plays a part in any asymmetric cell division, and also if phosphorylation plays a key part to the localization in these bugs.

References:

Alignments, B. S. L. M. S. (2001). BCM Search Launcher: Multiple Sequence Alignments. **http://searchlauncher.bcm.tmc.edu:9331/multi-align/multi-align.html Dec. 1 2001**.

Altschul, S. F. e. a. (1990). "Basic Local Alignment Search Tool." Journal of Molecular Biology **215**: 403-410.

Altschul, S. F. e. a. (1997). "Gapped Blast and PSI-Blast: a new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389-3402.

Boxshade3.21 (2001). Boxshade: Pretty Printing and Shading of Multiple-Alignment files. **http://www.ch.embnet.org/software/BOX_form.html Dec. 8 2001**.

Hecht, G. B. e. a. (1995). "An essential single domain response regulator required for normal cell division and differentiation in Caulobacter crescentus." EMBO J. **14**(16): 3915-3924.

Jacobs, C. e. a. (2001). "Dynamic localization of a cytoplasmic signal transduction response regulator controls morphogenesis during the Caulobacter cell cycle." Proc. Natl. Acad. Sci. **98**(7): 4095-100.

Jones, D. T. (1999). "Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices." Journal of Molecular Biology **192**: 195-202.

Lee, S.-Y. e. a. (2001). "Crystal structure of an activated response regulator bound to its target." Nat. Struct. Biol. **8**: 52-56.

Luscombe, N. e. a. (2001). "What is bioinformatics? A proposed definition and overview of the field." Method Inform Med **40**: 346-348.

NCBI-BLAST (2001). BLAST. **http://www.ncbi.nlm.nih.gov/BLAST/ Dec. 1 2001**.

PSIPRED (2001). PSIPRED Protein Structure Prediction Server. **http://bioinf.cs.ucl.ac.uk/psipred/ Dec. 8 2001**.

Scranton-WIT (2001). Similarity Search. **http://wit-scranton.mbi.scranton.edu/WIT2/CGI/sim_search.cgi?user= Dec. 1 2001**.

Thompson, J. D., et al (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Research **22**: 4673-4680.

TIGER (2001). TIGR BLAST Search Engine for Unfinished Microbial Genomes. **http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi? Dec. 1 2001**.