

Math/Stats 548, Winter 2004:
Computations in Biological Sequence Analysis

Laboratory Worksheet, Tuesday, Jan. 13.

I. Rooms and times. We will be switching the lab time to Fridays, 10-11 AM, unless I hear an emergency plea in the next 24 hours (i.e., by lecture tomorrow morning). The new room will be B743, a few doors down the hall. We also have available to us the Bioinformatics Computing Core (BiCC) or BI computer lab, 5631 Med Sci II. These two facilities are on two different local networks, and because of limits of site licences, they are not equivalent. BiCC has more specialized stuff about BLAST, but the math labs have licences, not surprisingly, for all the basic computational packages (Mathematica, Matlab, Maple, etc.). Both have various basic BI tools and modules, though I have to check up on the math labs to make sure the requests sometimes haven't expired. This usually applies to Perl modules and molecular visualization programs, such as Rasmol.

Since these are two local networks, you will have to set up two different accounts. This is relatively easy for the math labs, since there are a lot of people around to assist you, even after hours, and because we will set those accounts up in class. (You probably have already done this, if you are reading this document.) We will discuss in class about how to get the accounts in the BiCC set up.

II. 548 Lab Worksheets. This is the webpage which led you to this document: <http://www.math.lsa.umich.edu/dburns/547/548labworksheets.html>.

Ideally, meaning assuming I get the necessary time, there will be a worksheet for each lab session detailing what is expected for that day. This usually means "for that week" since, like most labs, we get things started but often not finished in the one hour. This may change this term, depending on your computing skills. That is, we may make the lab more independent in its workings, setting you off to work in teams on longer term projects, in which case the labs would become team meetings primarily to compare notes on the progress of your projects. That is for later, however.

There is also a 548 resource directory which should be accessible from the main page sidebar. The URL is <http://www.math.lsa.umich.edu/dburns/548/>.

III. Perl. Perl is a high level scripting language which is widely used for data handling in this field. It is relatively easy to get started in it. If you are unfamiliar with it, there is a tutorial available at http://www.math.lsa.umich.edu/dburns/548/perl_tut2003/. This is a tutorial in four lessons reproduced from last summer's BI shortcourses offered here at UM. There is also a zipped file of these lectures for convenient download to your personal computing directory: `perlfreight.tar.gz`. There are also annotated, illustrative examples of Perl routines from the book *Beginning Perl for Bioinformatics*, by James Tisdall (O'Reilly). This is a well-written book, and recommended if you have little computing experience. If you have experience, you may find *Learning Perl* by Schwartz and Phoenix (O'Reilly) more to your taste and speed. The illustration scripts are in `BeginPerlBioinfo.pm`, a ready to use module, or file of scripts.

While gearing up for our Perl exercises, we will need to streamline programs by using sub-routines. There is an open library of Perl scripts, and you should look at the archive of Perl

modules available at the **Comprehensive Perl Archive Network** (<http://www.cpan.org/>). Look especially at the BioPerl package of modules. The web resources page (<http://www.math.lsa.umich>) now has a link to the search page for CPAN. You should take a couple of minutes to explore this feature.

The 548 resources directory also contains a directory `bioperl-1.2/` containing these files. I will zip these up for convenience later today. The file name will be “`bperlfreight.tar.gz`”.

We will have to discuss whether separate meeting times will be required for a startup in Perl, or any of the other languages or programs we may come across during the term.

IV. Background to an Exercise. There are simple but illustrative calculations behind a paper of Karlin and Mrázek on predicting highly expressed genes. The paper is available from the 548 resource directory, where it is listed under `KMjbac.pdf`. You should begin having a look at this paper, especially the first two or three pages.

V. Related Seminars. There are, by chance, two seminars back-to-back in the math department Wednesday (tomorrow) which are relevant to the topics of this class. If you have a more quantitative background, you may find them of interest. Here are the abstracts.

Mathematical Biology Seminar
Wednesday January 14, 3:10-4:00pm, 4088 East Hall
H.T. Banks (North Carolina State University)

”Multiscale Issues and Inverse Problems in Computational System Biology”

A major intellectual challenge facing the research community entails organization and use of massive information at the gene/molecular (micro) level to understand system response at the organism (macro) level in biological systems. A number of fundamental issues (e.g., how to model uncertainty/variability in heterogeneous materials) in multiscale modeling and control are important in addressing this challenge. In this presentation we discuss some of these issues in the context of internal dynamics (molecular level) in system response models (organism level). We illustrate needs and ideas via results from specific examples: (i) examples involving molecular based reptation models for system level hysteresis in long molecular chain materials (e.g., polymers, tissues); (ii) examples from electromagnetics involving multiple dielectric polarization mechanisms in complex materials

Differential Equations Seminar
Wednesday January 14, 4:10-5:00pm, 3096 East Hall
Gilad Lerman (Courant Institute)

”Identifying Differentially Expressed Genes via Multiscale Geometric Analysis”

We confront some problems of data analysis (in particular outlier detection with applications to bioinformatics) and use ideas developed in harmonic analysis (specifically stopping-time constructions and multiscale geometric analysis). The first problem is the identification of differentially expressed genes or, more generally, the nonparametric detection of outliers in heteroskedastic data. We begin by assuming the data is normalized so that it is concentrated around a line. We suggest a multiscale construction for a ”strip” with varying width around the line. The strip is intended to separate the ”deviating”

points or genes from the rest. We may generalize our methods to the case where the data is not normalized a priori, so that we construct a strip of varying width around a curve. We also discuss briefly more general constructions, possible applications, difficulties, and relevant geometric information theory developed by Peter Jones and the speaker. This is a joint work with Joe McQuown and Bud Mishra as well as continuing theoretical work with Peter Jones. Remark: Despite the nature of the seminar no differential equations will be mentioned in the talk. However, we apply fundamental techniques of harmonic analysis arising in the study of differential equations.