Melisa Ricketts
12 December 2002
MB&B 452a
Final Project

Approaches to Vaccine Development Using Bioinformatics

Vaccine development has traditionally been a lengthy process. Animal studies with specific pathogens are used to find proteins that may interact with the immune system so that these individual proteins can be studied (Chakravarti et al., 2001; Zagursky and Russell, 2001). A great deal of time and expense is required, while often only a few potential candidate proteins for further evaluation are identified (Chakravarti et al., 2001; Zagursky and Russell, 2001). Bioinformatics provides a means of streamlining this process by reducing the number of potential targets to be investigated prior to beginning experiments. Additionally, the types of analyses performed may detect proteins that may not have been identified by other experimental assays. The preliminary studies using animals and microbial pathogens to detect potential proteins can be avoided, reducing the cost and time required to evaluate potential vaccine proteins.

The search for vaccine target proteins begins with determining the DNA sequence of the microbe of interest. At least 20 pathogenic bacteria have currently been sequenced (Jungblut, 2001). For a known sequence, open reading frames (ORFs) are predicted using a variety of programs. Examples of these programs include GeneMark, GLIMMER, ORPHEUS and Markov models (Zagursky and Russell, 2001). Each program incorporates information regarding codons, and applies it to the sequence to search for potential ORFs, or gene encoding regions (Zagursky and Russell, 2001).

Homology searches are carried out on the predicted ORFs in an attempt to determine function and other information about the potential gene and its protein product (Zagursky and

Russell, 2001). FASTA, BLAST, and variations of BLAST are commonly used algorithms for sequence comparison (Zagursky and Russell, 2001). FASTA generates tables of short query sequence to compare to the database, and BLAST expands upon short matches, to find the best scoring matches for the query sequence (Gerstein, 1999). The results of these searches yield preliminary information for use in other analyses, such as the function of proteins with homologous sequence, or even cellular localization.

A search for potential candidate proteins from *Neisseria meningitidis* serogroup B, an organism for which vaccine development has proven difficult, involved the methods described above (Pizza et al., 2000). The investigators sequenced this bacterium, and identified 2158 potential ORFs (Tettelin et al., 2000). After conducting homology searches, those predicted ORFs determined to likely be involved in functions in the cytoplasm were removed from the set of predicted ORFs, leaving only 570 for use in other analyses (Pizza et al., 2000). At this point, having narrowed the targets to a feasible number for experimentation, they began lab work.

Depending upon the type of immune response desired proteins that are expressed on the surface of a pathogen can be very important (Chakravarti et al., 2001; Ross et al., 2001; Zagursky and Russell, 2001). The previous example showed that homology searches can be used to determine the potential location of a protein within a cell, or pathogen. However, homologous genes are not always available in databases, and other methods of determining the cellular localization of unknown proteins are needed. The subcellular localization of potential proteins can be predicted using programs such as PSORT, ALOM, Pfam HMM, and ProtLock (Zagursky and Russell, 2001; Ross et al., 2001). These programs have different underlying methods, ranging from finding signal sequences or transmembrane segments, to looking at the amino acid content for making their predictions (Zagursky and Russell, 2001). The results from

any of these programs are important because it allows investigators to focus only on those proteins expressed on the surface, when it is important for the desired immune response to a vaccine.

The paper describing the complete sequence of the genome of *Streptococcus pneumoniae* detailed how the function and potential localization of some proteins was determined (Tettelin et al., 2001; Wizemann et al., 2001). The methods were similar to those described above, encompassing ORF determination, and sequence comparison, followed by subcellular localization determination (Tettelin et al., 2001; Wizemann et al., 2001). Only 130 of the 2687 predicted ORF were presumed to be expressed on the surface (Wizemann et al., 2001). Additional testing looking for vaccine targets was done on this small number of carefully selected potential proteins (Wizemann et al., 2001).

Gene expression analyses are also useful for vaccine development. Microarray experiments can be used to detect genes or clusters of genes that may make good targets by looking at differential expression patterns (Dhiman et al., 2002). Because the size of microbial genomes is generally small, it is often possible to spot the entire genome onto a single chip for experiments (Dhiman et al., 2002). Microarray results indicate when there are changes in expression of certain genes under specific conditions, including responses within a host cell or within the pathogen, to infection (Dhiman et al., 2002). Looking at the host response to an infection can identify intervention targets, and can even detect proteins that bind to MHC, an important component in certain immune responses (Dhiman et al., 2002). Potential vaccine candidates can also be tested for effectiveness using this method, prior to beginning additional trials.

Another approach to searching sequence for potential targets involves threading. Using coordinates of the structure of the binding cleft of MHC, all possible peptides from the

pathogen's protein sequence are assessed for the ability to interact with the cleft, and ranked accordingly (Altuvia et al., 1995). For this method, however, crystallographic data for the cleft of the MHC molecule of interest must be obtained (Altuvia et al., 1995). Crystallography allows researchers to determine the coordinates for use in the threading analyses (Altuvia et al., 1995). Ranks are given based upon interaction energies and contact potentials (Altuvia et al., 1995). Those pathogen peptide sequences that rank highest have the best probability of being interacting with the MHC molecules, which may include presentation on the surface of the host cell to immune system cells (Altuvia et al., 1995). For immune responses involving MHC antigen presentation, those pathogen peptides that will be processed and presented are important for appropriate interaction with the immune system to generate a strong response.

Microbial proteins that stimulate the strongest, appropriate immune response make the most effective vaccines, whether through contact of antibodies with surface proteins, or via processing and presentation by MHC to immune system cells. Finding these proteins is a long and difficult process, resulting in the lack of vaccines for many important microbial pathogens. Bioinformatics offers the ability to reduce vaccine development time and cost through the use of computer-based methods, prior to experimentation. Knowing the DNA sequence, allows one to predict ORFs and begin annotating proteins, including subcellular location, making the selection of proteins for evaluation more specific. The methods described here are only a few of the many ways that bioinformatics can be used to enhance vaccine development. These approaches should become standard in the search for vaccine targets, which would allow for faster results toward the prevention of illness.

References

Altuvia, Y., Scheuler, O. and Margalit, H. 1995. Ranking potential binding peptides to MHC molecules by a computational threading approach. *Journal of Molecular Biology*, 249: 244-250.

Chakravarti, D.N., Fiske, M.J., Fletcher, L.D. and Zagursky, R.J. 2001. Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. *Vaccine,* 19: 601-612.

Dhiman, N., Bonilla, R., O'Kane, D. and Poland, G.A. 2002. Gene expression microarrays: a 21st century tool for directed vaccine design. *Vaccine,* 20: 22-30.

Gerstein, M. 1999. Bioinformatics lecture notes for MB&B 452a. Yale University.

Jungblut, P.R. 2001. Proteome analysis of bacterial pathogens. *Microbes and Infection,* 3: 831-840.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Arico, B., Comanducci, M., Jennings, G.T., Baldi, L., Bartolini, E., Capecchi, B., Galeotti, C.L., Luzzi, E., Manetti, R., Marchetti, E., Mora, M., Nuti, S., Ratti, G., Santini, L., Savino, S., Scarselli, M., Storni, E., Zuo, P., Broeker, M., Hundt, E., Knapp, B., Blair, E., Mason, T., Tettelin, H., Hood, D.W., Jeffried, A.C., Saunders, N.J., Granoff, D.M., Venter, J.C., Moxon, E.R., Grandi, G. and Rappuoli, R. 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science,* 287: 1816-1820.

Ross, B.C., Czajkowski, L., Hocking, D., Margetts, M., Webb, E., Rothel, L., Patterson, M., Aguis, C., Camuglia, S., Reynolds, E., Littlejohn, T., Gaeta, B., Ng, A., Kuczek, E.S., Mattick, J.S., Gearing, D. and Barr, I.G. 2001. Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis. Vaccine,* 19: 4135-4142.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., Nelson, W.C., Gwinn, M.L., DeBoy, R., Peterson, J.D., Hickey, E.K., Haft, D.H., Salzberg, S.L., White, O., Fleischmann, R.D., Dougherty, B.A., Mason, T., Ciecko, A., Parksey, D.S., Blair, E., Cittone, H., Clark, E.B., Cotton, M.D., Utterback, T.R., Khouri, H., Qin, H., Vamathevan, J., Gill, J., Scarlato, V., Masignani, V., Pizza, M., Grandi, G., Sun, L., Smith, H.O., Fraser, C.M., Moxon, E.R., Rappuoli, R. and Venter, J.C. 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, 287: 1809-1815

Tettelin, H., Nelson, K.E., Paulsen, I.T., Eisen, J.A., Read, T.D., Peterson, S., Heidelberg, J., DeBoy, R.T., Haft, D.H., Dodson, R.J., Durkin, A.S., Gwinn, M., Kolonay, J.F., Nelson, W.C., Peterson, J.D., Umayam, L.A., White, O., Salzberg, S.L., Lewis, M.R., Radune, D., Holtzapple, E., Khouri, H., Wolf, A.M., Utterback, T.R., Hansen, C.L., McDonald, L.A., Feldblyum, T.V., Angiuoli, S., Dickinson, T., Hickey, E.K., Holt, I.E., Loftus, B.J., Yang, F., Smith, H.O., Venter, J.C., Dougherty, B.A., Morrison, D.A., Hollingshead, S.K. and Fraser, C.M. 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae. Science,* 293: 489-506.

Wizemann, T.M., Heinrichs, J.H., Adamou, J.E., Erwin, A.L., Kunsch, C., Choi, G.H., Barash, S.C., Rosen, C.A., Masure, H.R., Tuomanen, E., Gayle, A., Brewah, Y.A., Walsh, W., Barren, P., Lathigra, R., Hanson, M., Langermann, S., Johnson, S. and Koenig, S. 2001. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infection and Immunity*, 69(3): 1593-1598.

Zagursky, R.J. and Russel, D. 2001. Bioinformatics: use in bacterial vaccine discovery. *BioTechniques,* 31(3): 636-659.