Critical Analysis of the Computational Methods used to Discover Biomarkers to assist in the Early Detection of Disease

Vishnu Patankar
Biochem 218,
March 2005

1. Introduction

Using proteins as biomarkers has long been considered a promising clinical diagnostics approach for drug discovery and development. Some biomarkers, such as prostate-specific antigen, have been in use for many years. Many other potential biomarkers are being reported in the literature almost weekly, although few have been translated into the diagnostic arena. Progress has not been as rapid as we would like, especially given the advances in our understanding of the process by which disease develops and becomes lethal [45]. Bio-software has a pivotal role to play here - for e.g., to leverage knowledge gained from work with tumors where the biology is known and apply this to the complex proteomics of serum and other body fluids.

In this paper, we analyze the two most popular computer database search algorithms used in protein identification but will begin with a few preliminaries and motivation for the analysis.

1.1 Early Detection of Disease

What follows are study results from recent clinical and clinico-algorithmic studies relating to the value of biomarkers and the role played by computer algorithms in the early detection of disease.

1.1.1 Clinical study

Heart Disease: Levels of a specific protein biomarker in the blood could predict the risk of heart attack or death in those with coronary heart disease. The protein called placental growth factor (PGF) is known to trigger inflammation within hardened and narrowed coronary arteries. A recent study [43] suggests that PGF's presence could perhaps be used as a 'marker' for prognosis in heart disease. Levels of PGF were measured in a group of 547 patients with known heart disease. PGF was also measured in another group - of 626 patients presenting with acute chest pain in an emergency department. In those with heart disease, elevated PGF indicated an increased risk of heart attack or death within 30 days. In those with chest pain, raised PGF meant a three fold increased risk of heart attack or death. The study conclude that PGF is indeed a valuable biomarker for heart attack or heart death and that therapies targeting the inflammatory action of PGF would be a good approach for treating heart disease.

1.1.2 Clinico-Algorithmic studies

Prostrate cancer: The prostate-specific antigen test has been a major factor in increasing awareness and better patient management of prostate cancer (PCA), but its lack of specificity limits its use in diagnosis and makes for poor early detection of PCA. Identifying better biomarkers for early detection of PCA using protein profiling technologies can simultaneously resolve and analyze multiple proteins. Evaluating multiple proteins will be essential to establishing signature proteomic patterns that distinguish cancer from noncancer as well as identify all genetic subtypes of the cancer and their biological activity. One study [41] used a protein biochip surface enhanced laser desorption/ionization mass spectrometry approach coupled with an artificial intelligence learning

algorithm to differentiate PCA from noncancer cohorts. A blinded test set, separated from the training set by a stratified random sampling before the analysis, was used to determine the sensitivity and specificity of the classification system. A sensitivity of 83%, a specificity of 97%, and a positive predictive value of 96% for the study population and 91% for the general population were obtained when comparing the PCA versus noncancer (benign prostate hyperplasia/healthy men) groups.

Ovarian cancer: Another study [42] used proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary. A training set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analyzed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders. The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non-malignant disease, 63 were recognized as not cancer. This result yielded a sensitivity of 100% (95% CI 93--100), specificity of 95% (87--99), and positive predictive value of 94% (84--99). These findings justify a prospective population-based assessment of proteomic pattern technology as a screening tool for all stages of ovarian cancer in high-risk and general populations.

1.2 Proteomics

Proteomics is the systematic study of the many and diverse properties of proteins in a parallel manner with the aim of providing detailed descriptions of the structure, function and control of biological systems in health and disease. Advances in methods and technologies have catalyzed an expansion of the scope of biological studies from the reductionist biochemical analysis of single proteins to proteome-wide measurements. Proteomics and other complementary analysis methods are essential components of the emerging 'systems biology' approach that seeks to comprehensively describe biological systems through integration of diverse types of data and, in the future, to ultimately allow computational simulations of complex biological systems.

1.3 Protein Sequencing in relation to DNA Sequencing

Forward Genetics a key element of reductionist research approaches in the 1980s attempted to move from an observed phenotype or function to the relevant genes and their products that caused that phenotype.

Reverse Genetics benefited from the advent of large-scale sequencing projects and their results [1] catalyzing the development of *reverse* approaches, which attempted to move from the gene sequence to function and phenotype. Such approaches included the observation of clusters of mRNA species showing coordinated expression patterns in different cellular states, either by expression arrays or by serial analysis of gene expression (SAGE [2]).

The rapid identification of proteins was limited only by our capacity to extract sequence information from proteins and peptides, and to correlate this information with the sequence databases. Mass spectrometry and database search algorithms fill this gap.

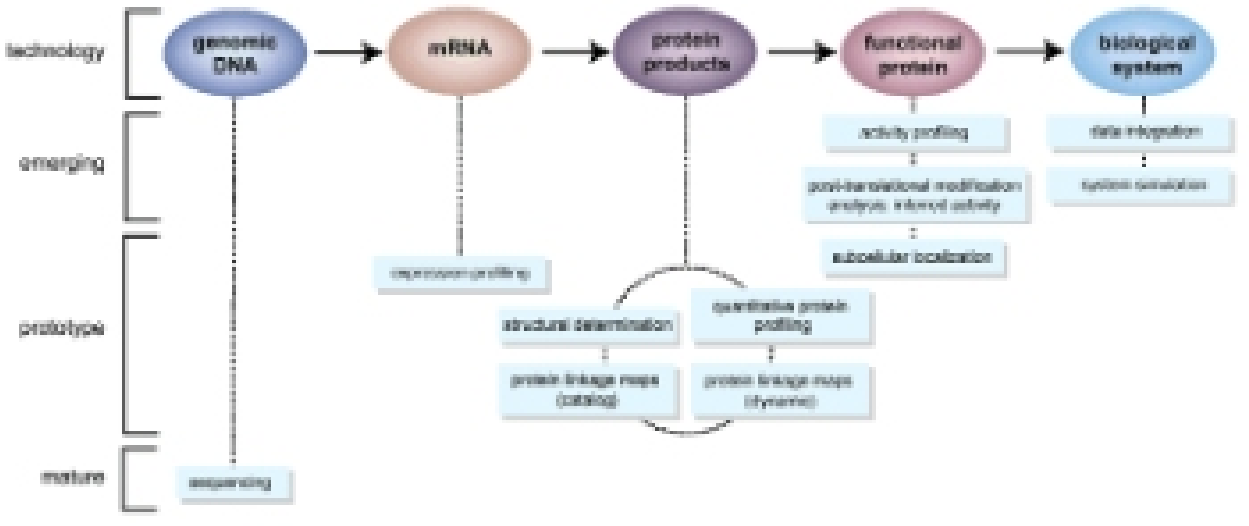

Figure 1: The current status of proteomic technologies.

The different data typically collected in proteomic research and the available technologies are listed. The relative maturity of the proteomic technologies and other key discovery science tools is apparent from the position of the respective technology on the graph.

2. Protein identification methodology

Broadly, two steps constitute the methodology used to identify proteins - Mass Spectrometry and Database Search. A protein mixture is digested, and the resulting peptides are analyzed by MS/MS to obtain experimental spectra. Search programs find database candidate sequences whose theoretical spectra are compared to the experimental spectrum. The best match (highest-scoring candidate sequence) defines the identified database peptide and the corresponding database protein. Validation software then determines whether the peptide and protein identifications are true or false.

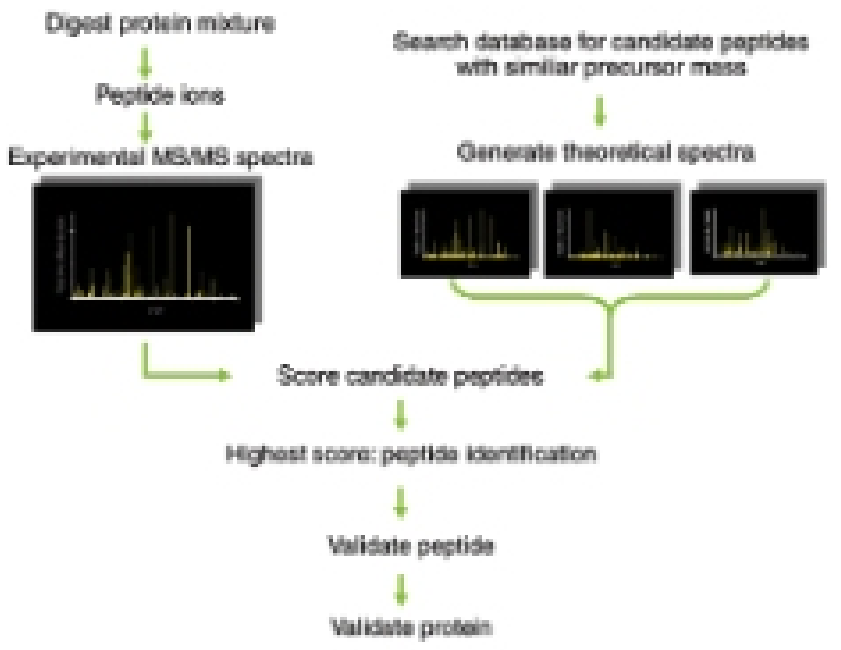

Figure 2. Overview of the protein identification process.

2.1 Mass spectrometry

A mass spectrometer measures the mass-to-charge ratio of charged species under vacuum and comprises an ionization source and a mass analyzer. In the late 1980s, two methods were developed that allowed the