

Scientists from the RAND Corporation have created this model to illustrate how a "home computer" could look like in the year 2004. However the needed technology will not be economically feasible for the average home. Also the scientists readily admit that the computer will require not yet invented technology to actually work, but 50 years from now scientific progress is expected to solve these problems. With teletype interface and the Fortran language, the computer will be easy to use.

Proteomics 2.0:
better, faster, cheaper



CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

Ron Beavis

Canada Research Chair in
Experimental Bioinformatics
University of British Columbia

1. Proteomics, as currently practiced, works surprisingly well.



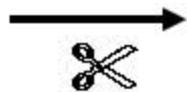
CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

Protein

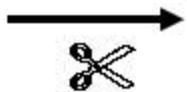
Proteolytic
Peptides

Peptide
Fragments

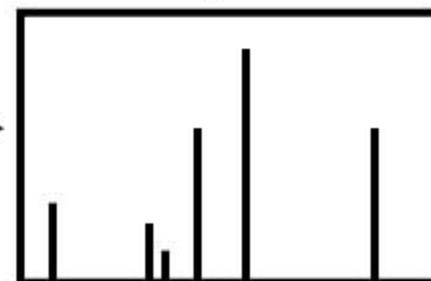
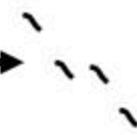
Fragment
Mass Spectrum



(enzymatic)



(gas phase)



LAB

CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

Basic experiment for protein identification

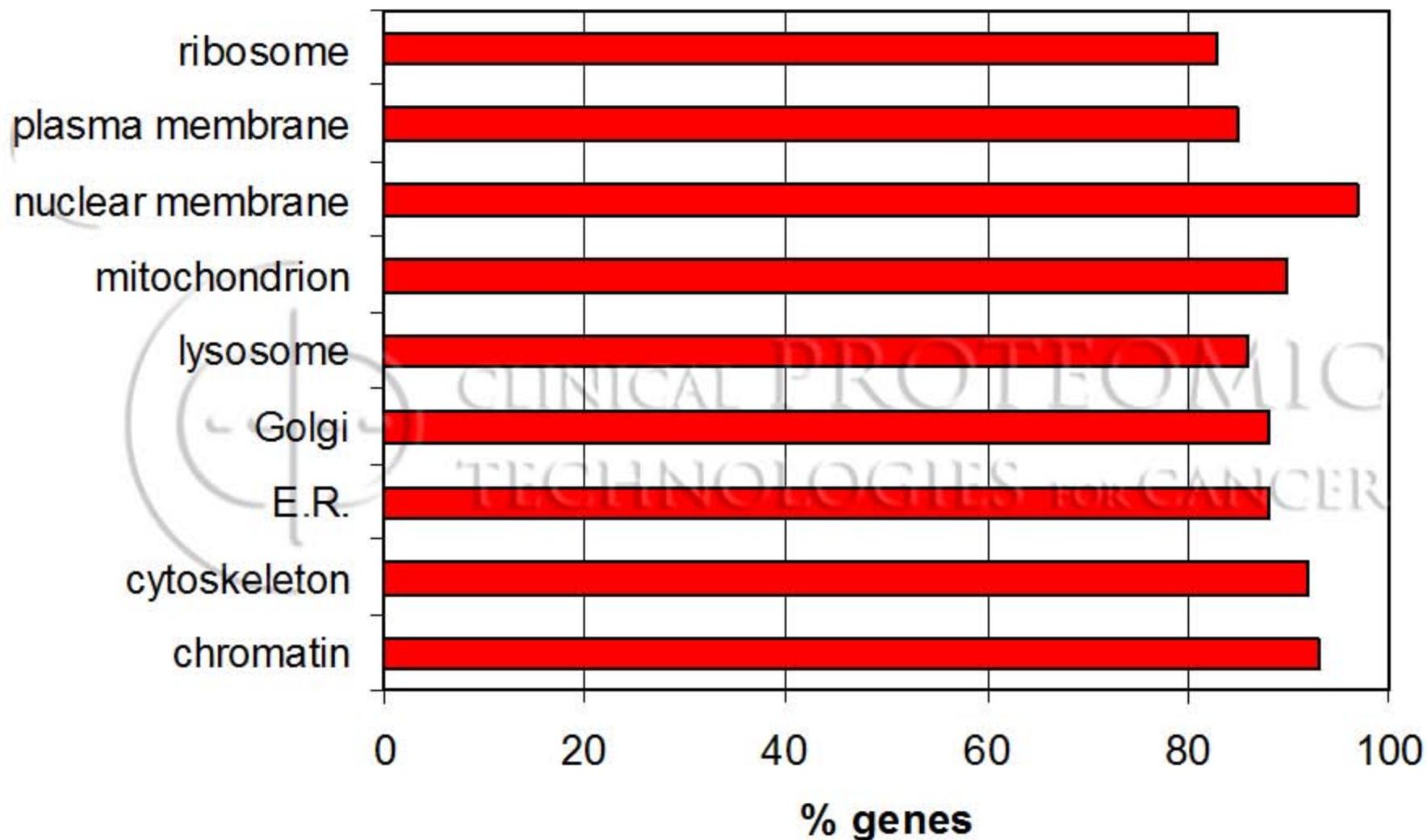
spectra

sequences

Generic search engine

Test all
cleavages,
modifications,
& mutations
for all sequences

Protein identification informatics



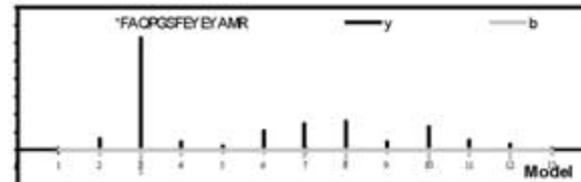
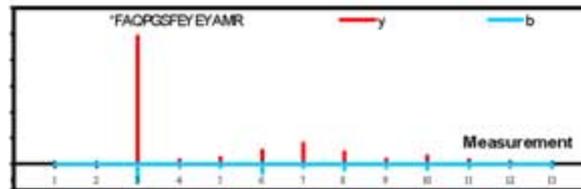
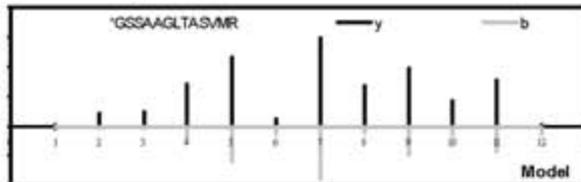
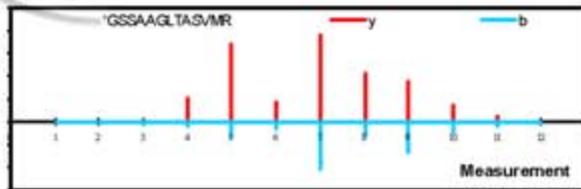
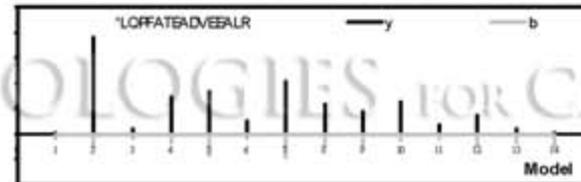
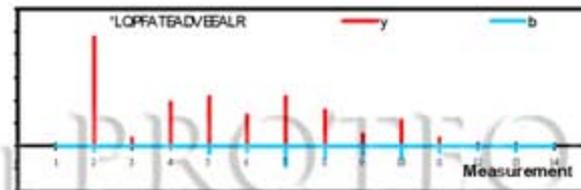
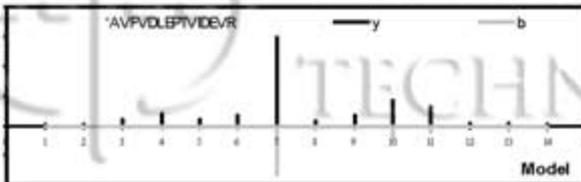
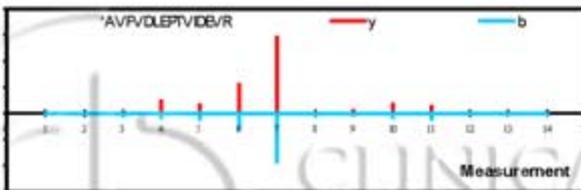
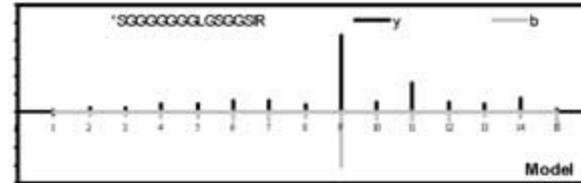
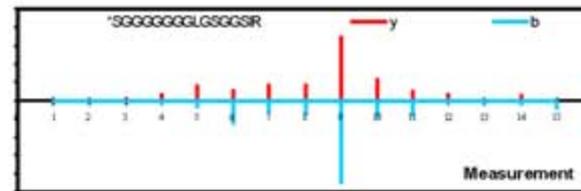
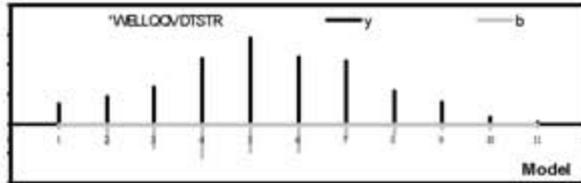
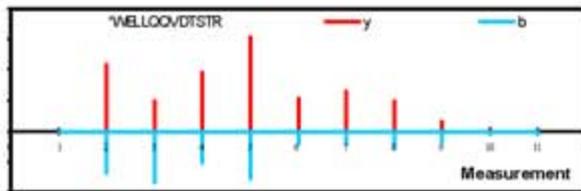
Human genes observed: GO components



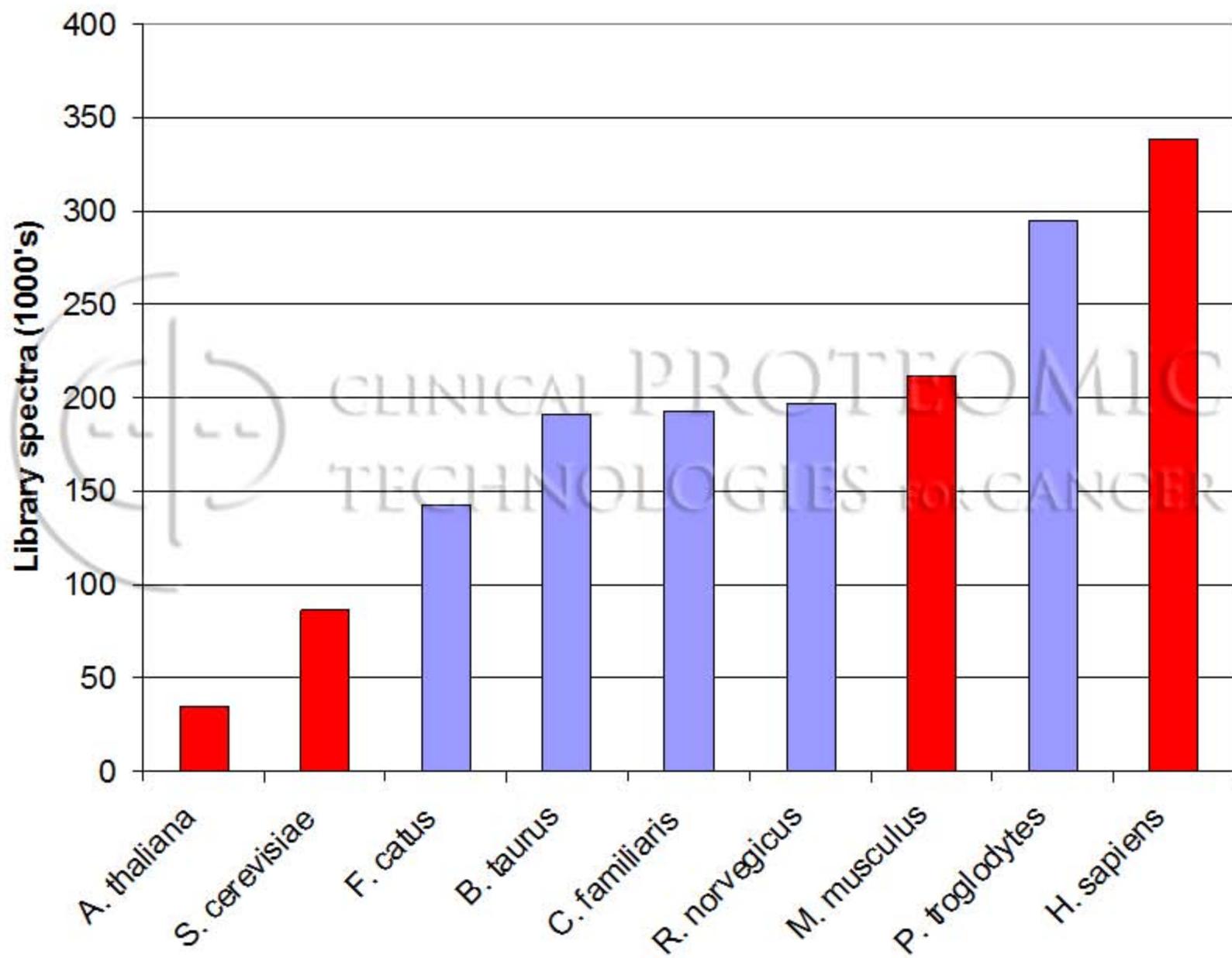
Usage pattern for a public proteomics resource

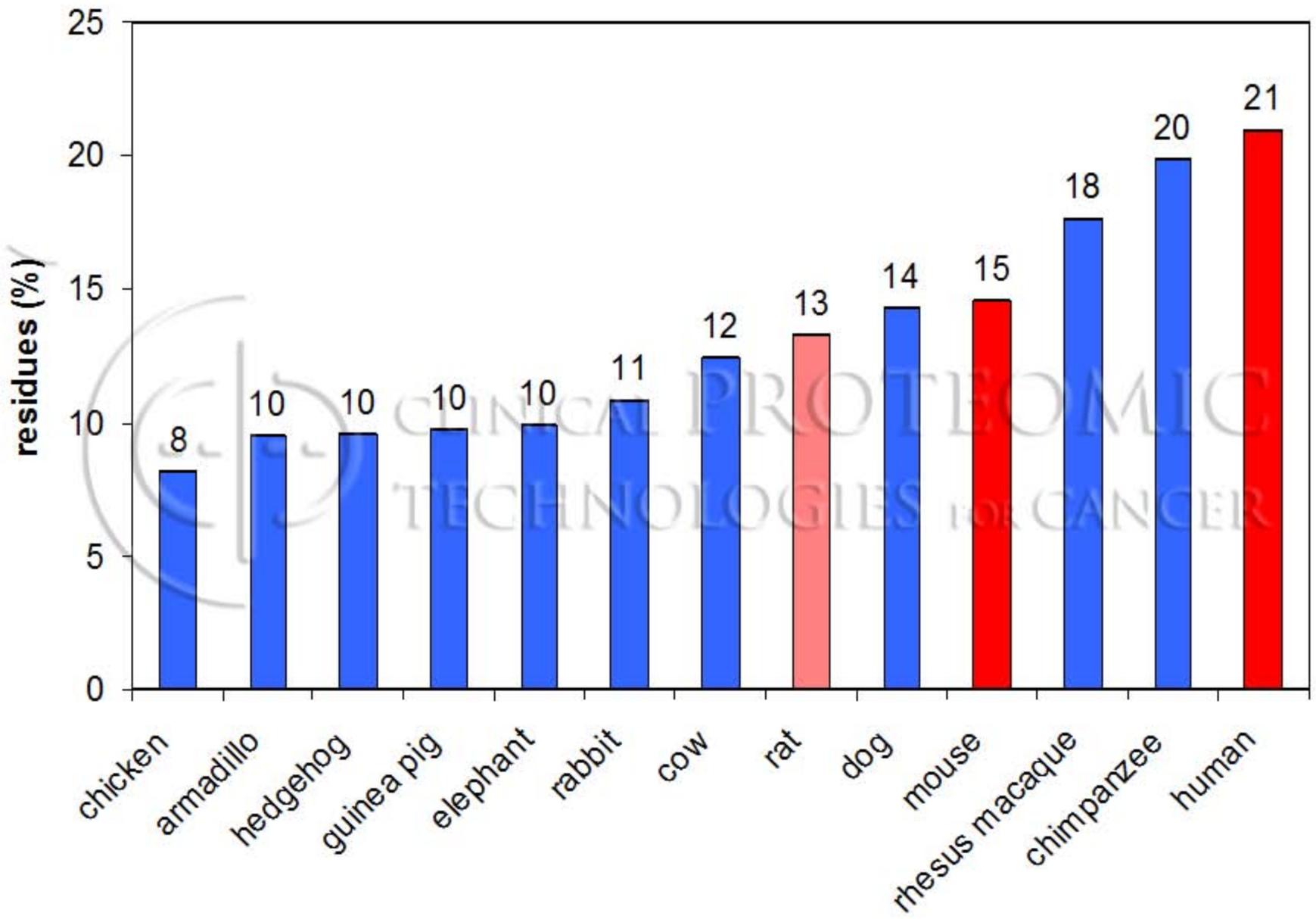


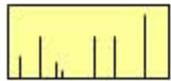
Most proteins show very reproducible peptide patterns



Examples of measured and modelled ion intensities for fragmentation



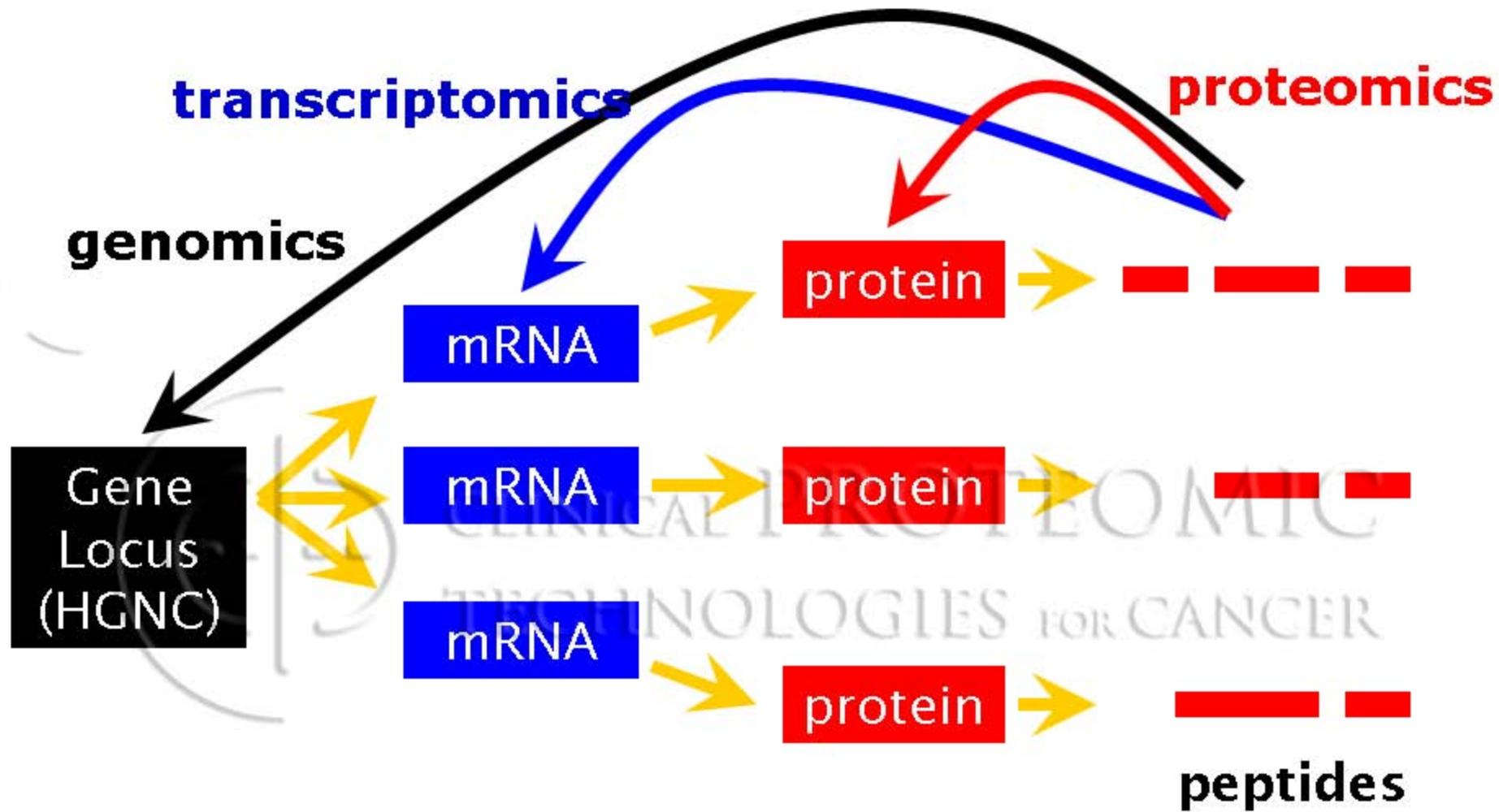




1. Array spectra in computer memory
2. Compare with observed data



Use the experimental spectra themselves



Organization of biological sequences

Things that haven't worked out so well



CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

1. Laboratory automation

It had been anticipated that proteomics would follow the same model as genome sequencing, where a small number of highly automated pipeline-style laboratories would produce all of the necessary proteomics information.

Laboratory information handling systems proved very difficult to write and customize.

2. Informatics Standardization

Based on the original aspirations of HUPO-PSI, there should be standard XML formats allowing the interchange of both raw and processed experimental data and conditions. The output for informatics analysis should also be in a consistent format.

Standards committees have proven to be too slow (and easily distracted) to generate file formats that effectively capture domain-specific information

3. Accession number rot

Protein accession numbers should provide a stable method of reporting protein identification experimental results. Publications have relied on the long-term maintenance of sequence databases for this purpose.

Examination of a 2003 paper on platelet proteomics showed that fully 1/3 of the accession numbers corresponded to protein sequences that were no longer available.

4. The identification numbers game.

In an effort to demonstrate the effectiveness of proteomics techniques, the quality of an analysis is often measured by the number of identifications, rather than the statistical significance of the results.

Competition with cDNA chip technologies and initial over optimism about the useful dynamic range of protein identification experiments.

Things that we will need over
the next 5 years



PRIDE (EBI) Repository of published identifications:
PRIDE XML

PeptideAtlas (ISB) Data and results oriented resources for the US scientific community
List of peptides known to have been observed.
mzXML, pepIdent

GPMDDB Bioinformatics resource for ID validation and research:
BIOML, mzData, mzXML, GAML

P41 - a relatively large, stable resource for algorithm development, implementation and data storage

Coordinated funding programs that reflect the lifecycle of computational efforts and provide a stable flow of innovation to the private sector

SBIR/STIR - the lesson of existing informatics in this area is that without the active participation of small-to-medium sized private sector partners, good ideas may not be widely accepted by the experimental community.

Proteomics 2.0

Improved handling and manipulation of large datasets
Improved integration between boutique proteomics databases
Proteomics informatics strategies:

Requires:

Requires more modern web technologies

Requires:

= ~~CD~~ Only available, aggressive lossy protein specific accession number compression strategies (MP3 for proteomics)

disambiguation

= ~~Conviction~~ ~~holoproteins~~ and demonstrated long stable, published query interfaces for public

data resources

results displays