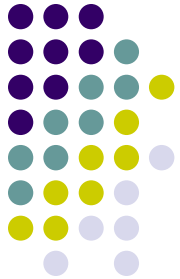
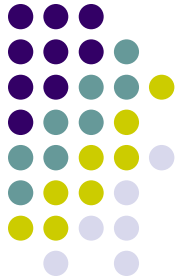
The Basics



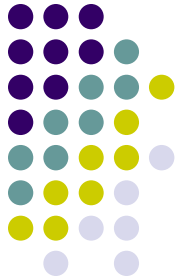
- Pattern Discovery And Recognition
  - A Single Biomarker For Ovarian Cancer Is Proving Elusive
    - CA125 As A Diagnostic Is Unreliable
    - Any Tumor Secreted Protein/Peptide Will Likely Have Low Concentration In Serum
  - Patterns Of Proteins/Peptides Reflect Systematic Response To Tumor Appearance
    - Hormonal Effects
    - Immunologic Response
    - Chaotic Changes
    - Multiple Patterns Per Disease State

The Basics



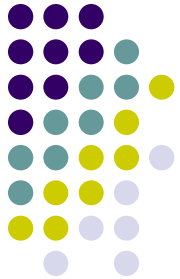
- Feature Selection
  - Treat Mass Spec Data As A Signal
  - For Ciphergen Mass Spec, Each M/Z Line Is A Feature
  - Signal Consists Of 15,280 Features
  - Finding Optimal Feature Set Is Overwhelming Using Conventional Methods
    - For Five Feature Pattern, There Are $15{,}280^5$ Combinations
    - Explicit Search Cannot Be Finished In A Lifetime

The Knowledge Discovery Engine™



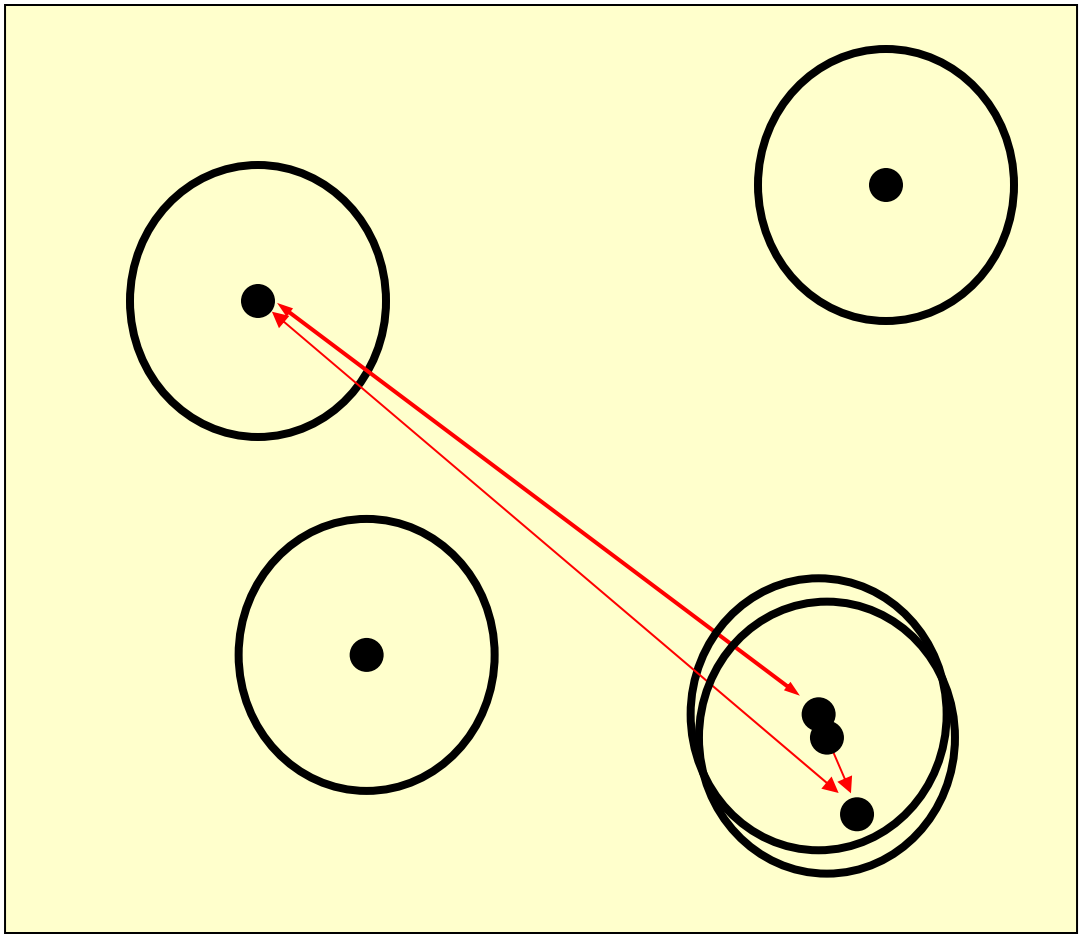
- Finds A Near Optimal Feature Set For Use In A Pattern Recognition Algorithm
- Three Components
  - A Genetic Algorithm Selects Features
  - A Self Organizing, Adaptive Pattern Recognition Algorithm Clusters Data
  - A Simple Statistic Provides Information On Cluster Homogeneity

Self Organizing Adaptive Pattern Recognition

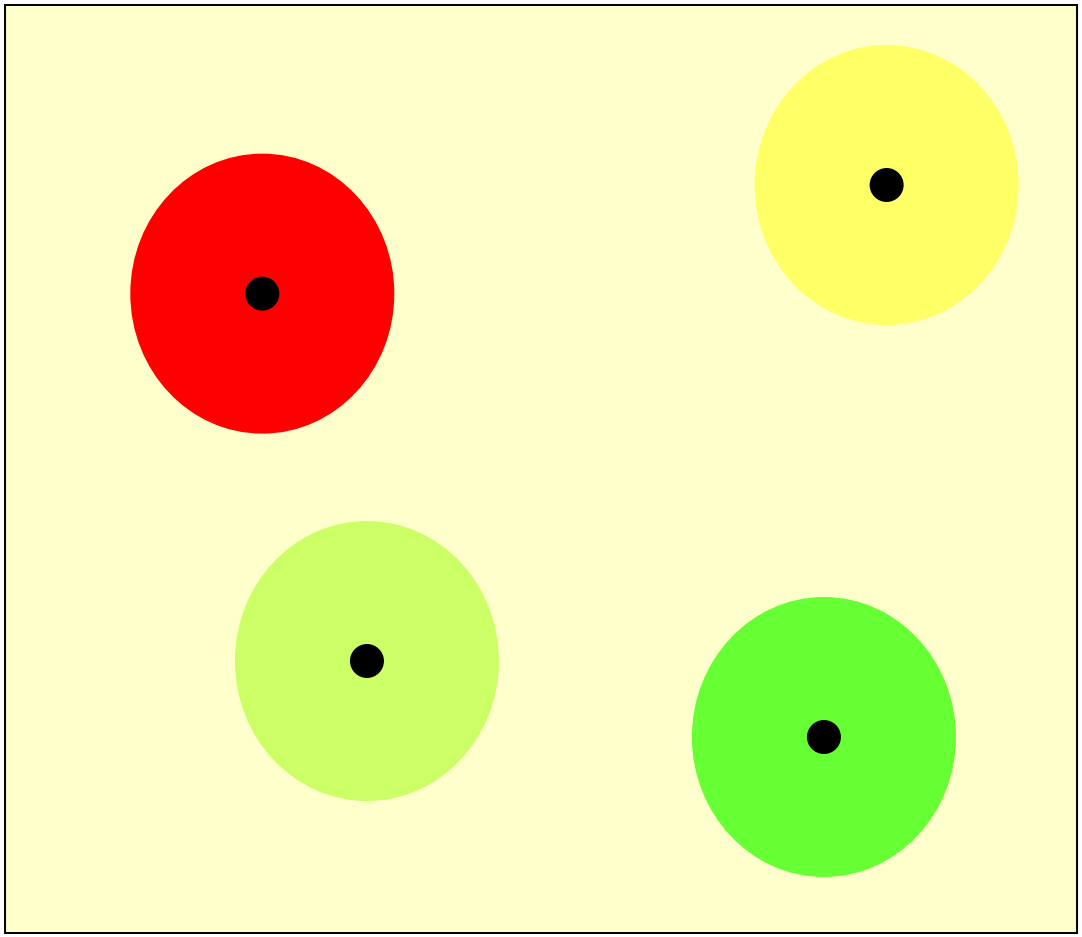


- Proteome Quest® Uses The Lead Cluster Map
  - *N*-Dimensional Euclidian Distance Based Classification
  - Adaptive - Always Learning
  - Vigilant - Recognizes Novel Event In Data Stream
  - Fast, One Pass Training

The Lead Cluster Map



How Homogenous Are The Clusters?

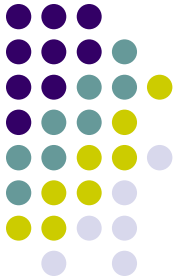


The Map Homogeneity Is Computed As The Average Homogeneity Of Each Cluster Across All The Clusters. This Is Used As The Fitness In The Genetic Algorithm

When The Map Is 100% Organized, The Resulting Clusters Are Evaluated For How Homogenous They Are With Respect To Biological State.

Green Is 0% Diseased
Yellow Green Is 25% Diseased
Yellow is 50% Diseased

The Genetic Algorithm



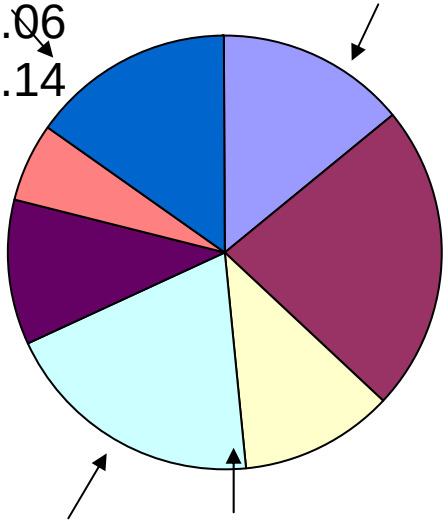
- Simulates Natural Evolution To Optimize Feature Set
- Operations Follow Biologic Adaptation
  - Probabilistic Fitness Selection
  - Mating (Crossover)
  - Reproduction
  - Mutation
  - Culling

The Genetic Algorithm

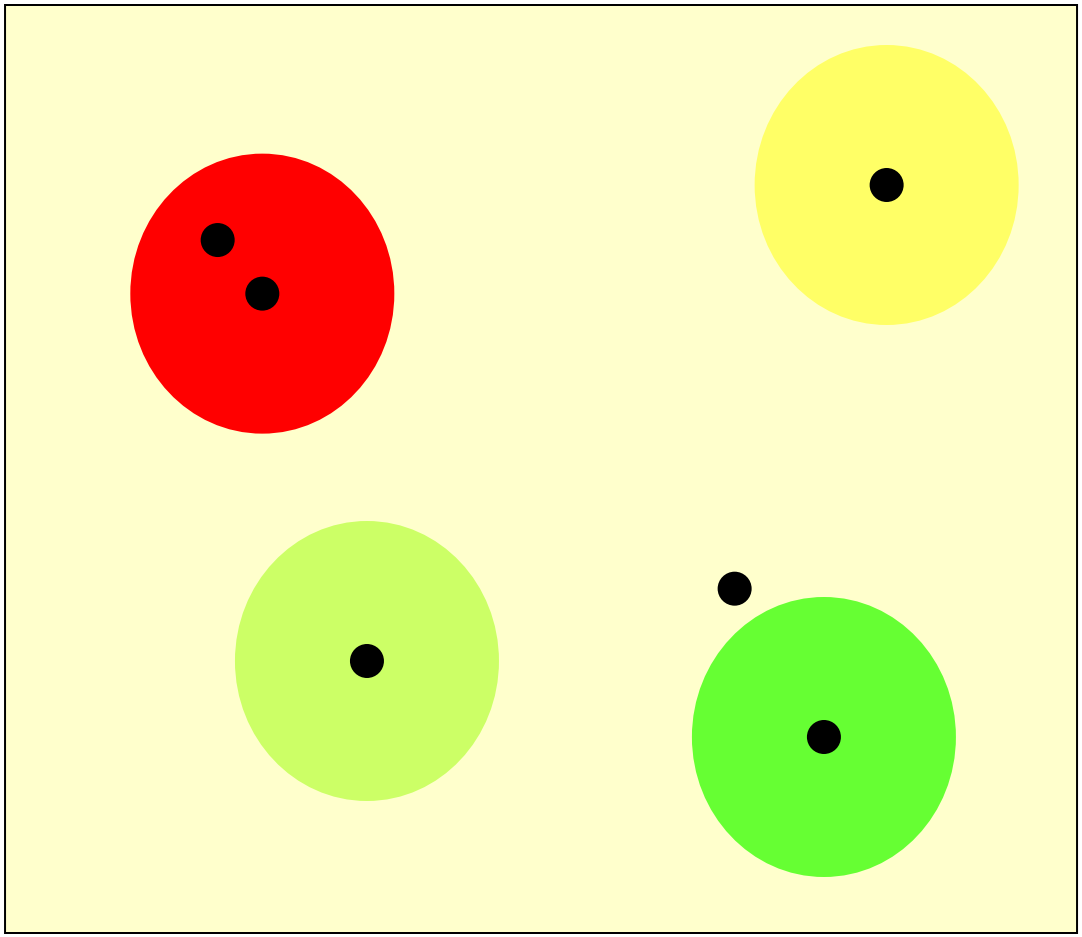
[8428,	2254,	6927,	9908,	2399]	0.81	0.22
[7769,	6156,	9001,	951,	646]	0.73	0.14
					3.61	

1.76

Be Duplicated



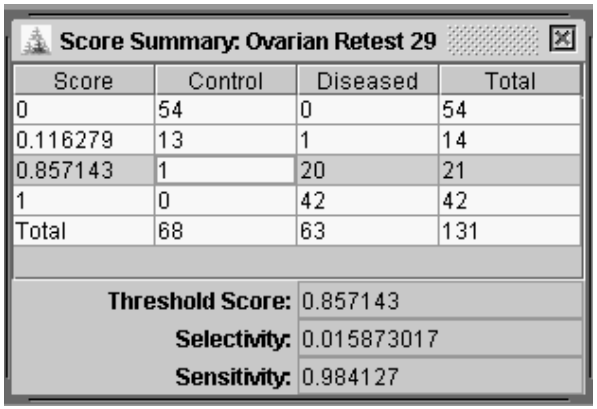
Diagnostic Model Use



When An Unknown Specimen Arrives In The Lab Its Protein Pattern Is Compared To The Diagnostic Model. The Specimen Is Scored Based On The Cluster It Falls Into. The Score Is 100% Diseased In This Example

If The Specimen Does Not Fall Into A Cluster, It Is Scored As The Nearest Cluster And Noted As Novel. This Sample Would Be Scored As 0% Diseased But New To The System!

Modeling Results For Ovarian Cancer



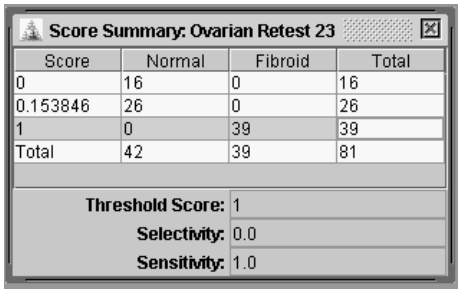
Score Summary: Ovarian Retest 29

Score	Control	Diseased	Total
0	54	0	54
0.116279	13	1	14
0.857143	1	20	21
1	0	42	42
Total	68	63	131

Threshold Score:	0.857143
Selectivity:	0.015873017
Sensitivity:	0.984127

Node	Count	State	StateSum	Error	349.44058	785.44338	245.53704	8749.6273	8003.3008
0	42	1	36	6	0.996845	0.216242	0.396389	0.063826	0.345897
1	43	0	5	5	0.995847	0.216443	0.597174	0.062445	0.408394
2	25	1	25	0	0.993629	0.18432	0.457702	0.0471922	0.210048
3	9	0	0	0	0.997733	0.426925	0.715582	0.0410085	0.336809
4	3	0	0	0	0.974219	0.377149	0.655183	0.0844548	0.53315
5	13	0	0	0	0.993606	0.202087	0.65194	0.0515213	0.265342
6	1	0	0	0	1	0.560264	0.539024	0.0747843	0.464684
7	4	0	0	0	0.846141	0.262706	0.876987	0.0368215	0.376691

Fibroid V Cancer



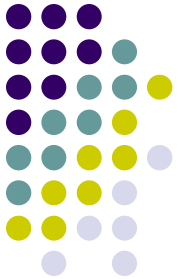
Score Summary: Ovarian Retest 23

Score	Normal	Fibroid	Total
0	16	0	16
0.153846	26	0	26
1	0	39	39
Total	42	39	81

Threshold Score:	1
Selectivity:	0.0
Sensitivity:	1.0

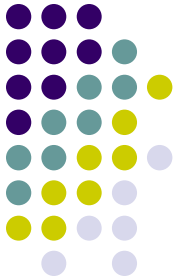
Node	Count	State	StateSum	Error	8688.6274	8118.9303	9175.448	364.60468	6877.1305	8672.9762	4609.5839	280.32307	244.66041	353.28911
0	40	1	40	0	0.0452824	0.193639	0.0815192	0.129139	0.0302258	0.0476951	0.0953376	0.304083	0.217213	0.461908
1	26	0	4	4	0.056271	0.24848	0.186583	0.133643	0.0360071	0.0584209	0.150795	0.349218	0.580895	0.393635
2	9	0	0	0	0.0545433	0.240165	0.202003	0.137641	0.0361038	0.0574982	0.176575	0.334897	0.373167	0.605806
3	3	0	0	0	0.0641974	0.183802	0.509919	0.122027	0.0468114	0.0680114	0.294685	0.321132	0.486563	0.531164
4	6	1	6	0	0.0374897	0.185343	0.0600801	0.120849	0.0248064	0.0392355	0.0693076	0.256004	0.547108	0.260906

KDE Advantages and Disadvatages



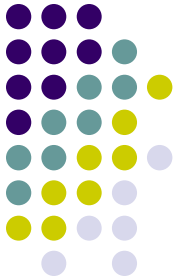
- Advantages
  - Efficient
  - Begins With No Assumptions
  - Intrinsically Non-Linear
- Disadvantages
  - Over-fitting
  - Sensitive to Artifacts

Key Features Of LCM Model



- Computationally Efficient – 2.0 Ghz Computers Can Process > 1,000,000 Samples A Day
- Ability To Gain Experience Is Built In
- Recognizes Novel Proteomic Patterns
  - New Disease Variants
  - Identify Pockets Of Emerging Disease
  - Provide General Monitoring Of Target Population

Pre-Analytic Issues



- Sample Set
  - Sample Collection
    - Tube Type
    - Subject Condition
  - Sample Preparation
    - Clotting Time
    - Time On Clot
    - Time To Freezer

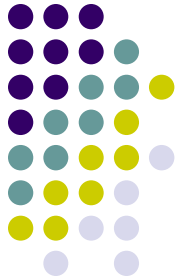
Pre-Analytic Issues

- Sample Characterization
  - Mislabeling
  - Corroboration of Histo-Path
  - True Control

Data Collection Issues

- Machine Variation
  - Day To Day
  - Site To Site
  - Duplicate Or Triplicates
    - Sequential
    - Random
  - True Validation

Post-Analytic Issues



- Role Of Variability
- Meaning Of Indeterminate Results
- Clinical Use *w.r.t.* Sensitivity and Specificity