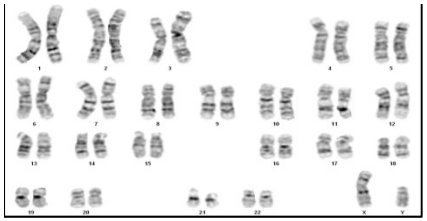
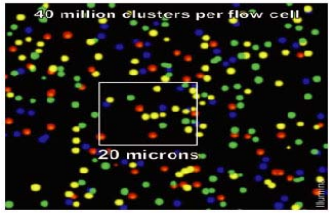
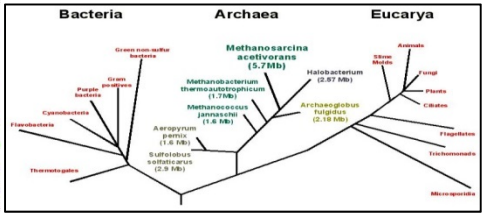


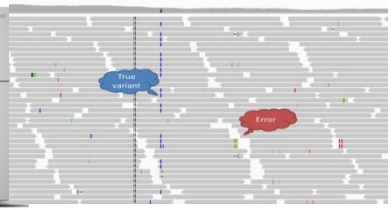
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA

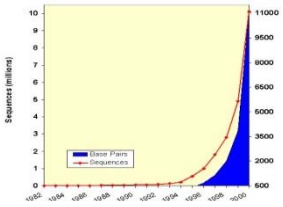


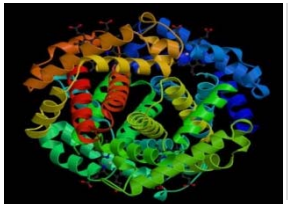


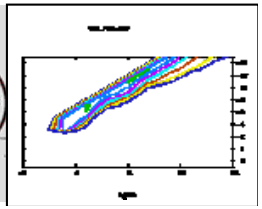
Computer Lab:
Feature selection and Cluster analysis

孟宇琦 （Meng Yuqi）
北京大学生物信息学中心
Center for Bioinformatics, Peking University

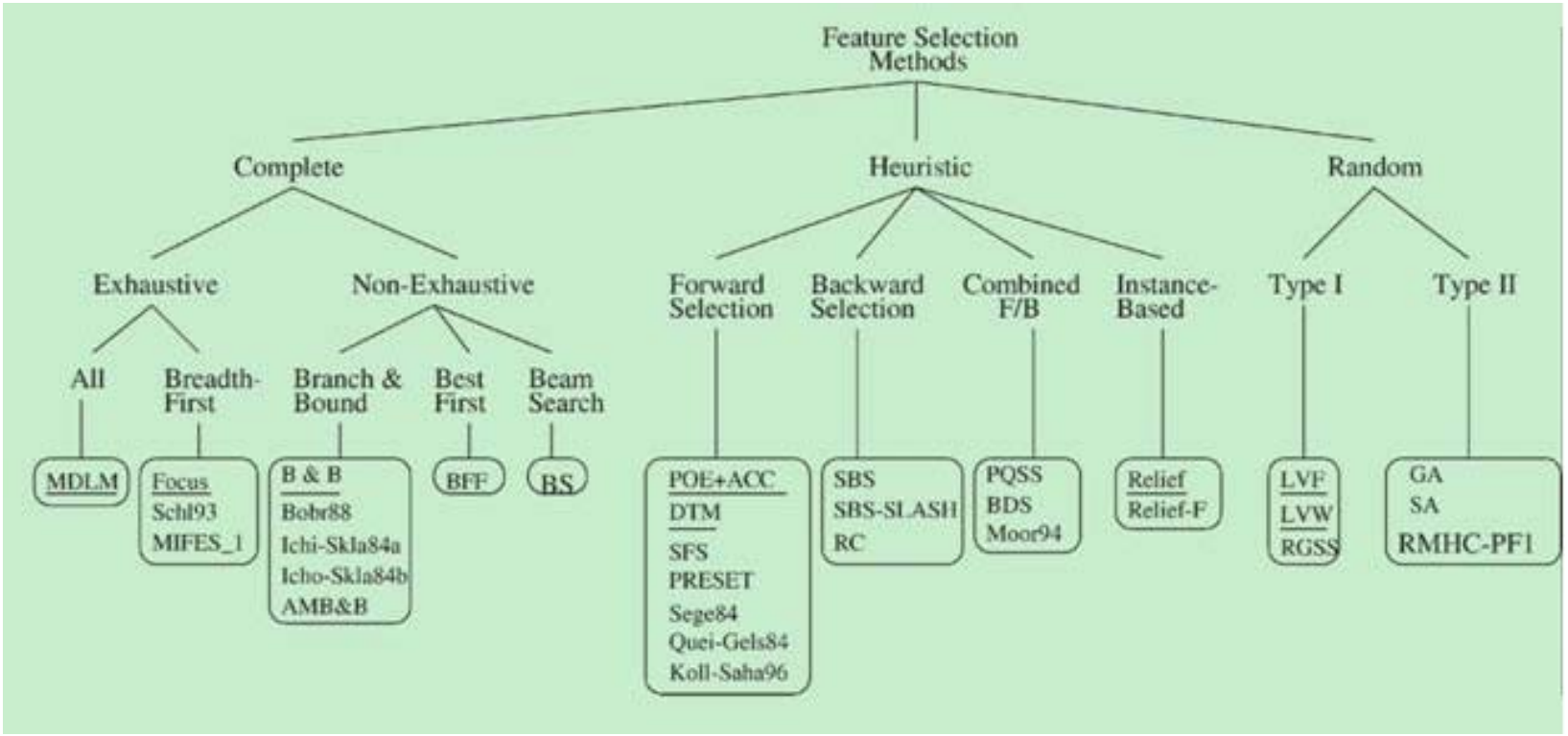








Find The Optimal Subset



The way to find the optimal subset (M. Dash and H. Liu 1997)

Introduction Of Heuristic Search

- **SFS , Sequential Forward Selection**

  Set of variables starts from an empty set, each time we select a variable to join the subset and the optimal solution in the evaluation is selected. Each time select a optimal variable to join, a simple greedy algorithm.
- **SBS , Sequential Backward Selection**

  Set of variables starts from an set which has all variables ,each time we remove a variable from the subset and the optimal solution in the evaluation is selected.
- **BDS , Bidirectional Search**

  Using a sequence forward selection (SFS) starts from the empty set, while using the sequence backward selection (SBS) to start the search from the universal set, when the two are the same, stop the search.

Introduction Of Heuristic Search

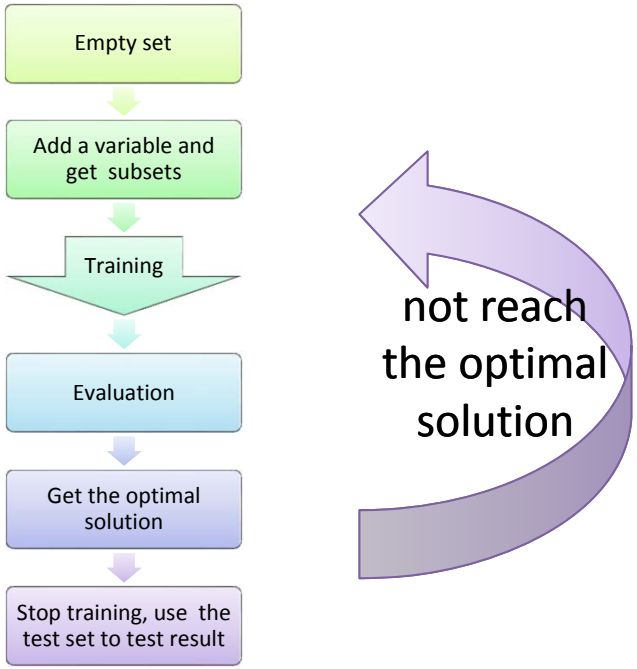
- **LRS , Plus-L Minus-R Selection**

  Starts from the empty set, each time join L variables, and then remove R variables, the optimal solution in the evaluation is selected.(L > R)

  Starts from the universal set, each time remove R variables, and then join L variables, the optimal solution in the evaluation is selected. (L < R)

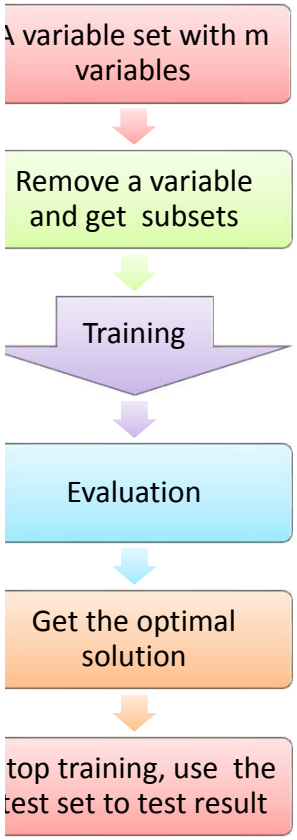
- **Sequential Floating Selection**

  Sequential Floating Selection is from the Plus-L Minus-R Selection , the differs is : the L and R is not fixed ,it will changing.



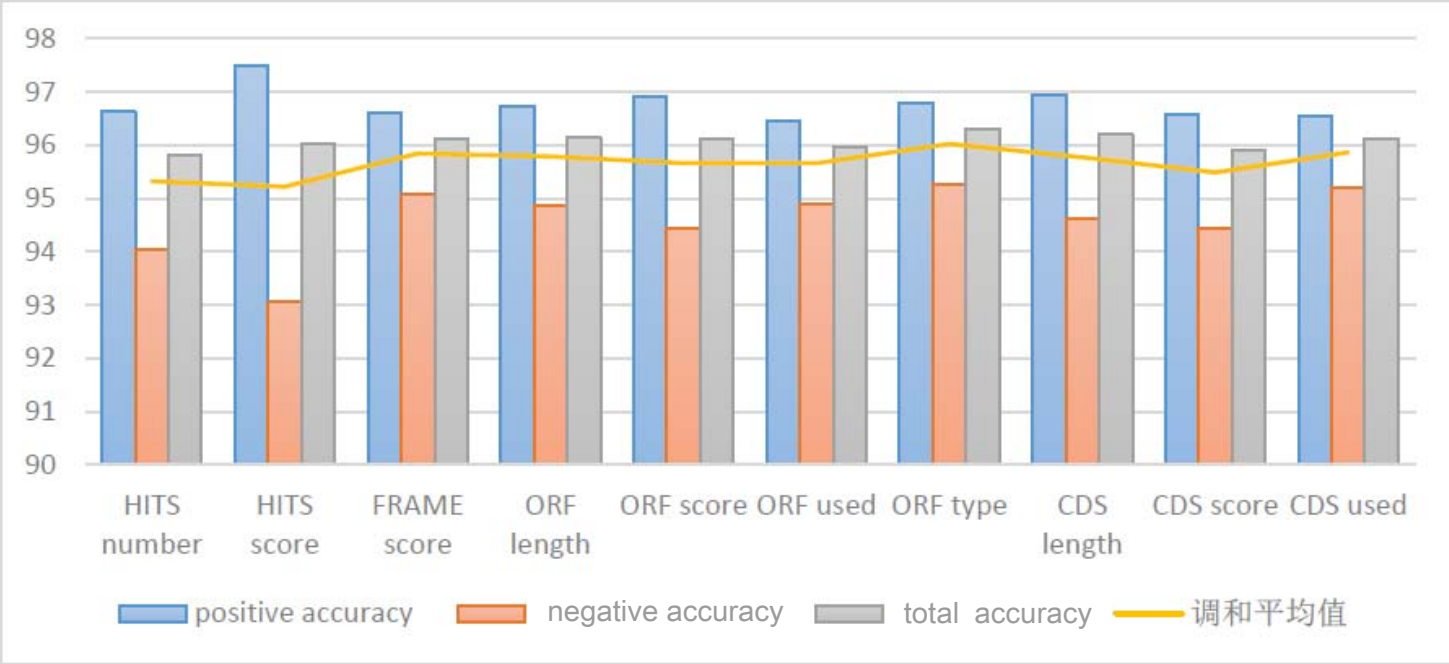
SFS

Set of variables starts from an empty set, each time we select a variable to join the subset and the optimal solution in the evaluation is selected. Each time select a optimal variable to join, a simple greedy algorithm.

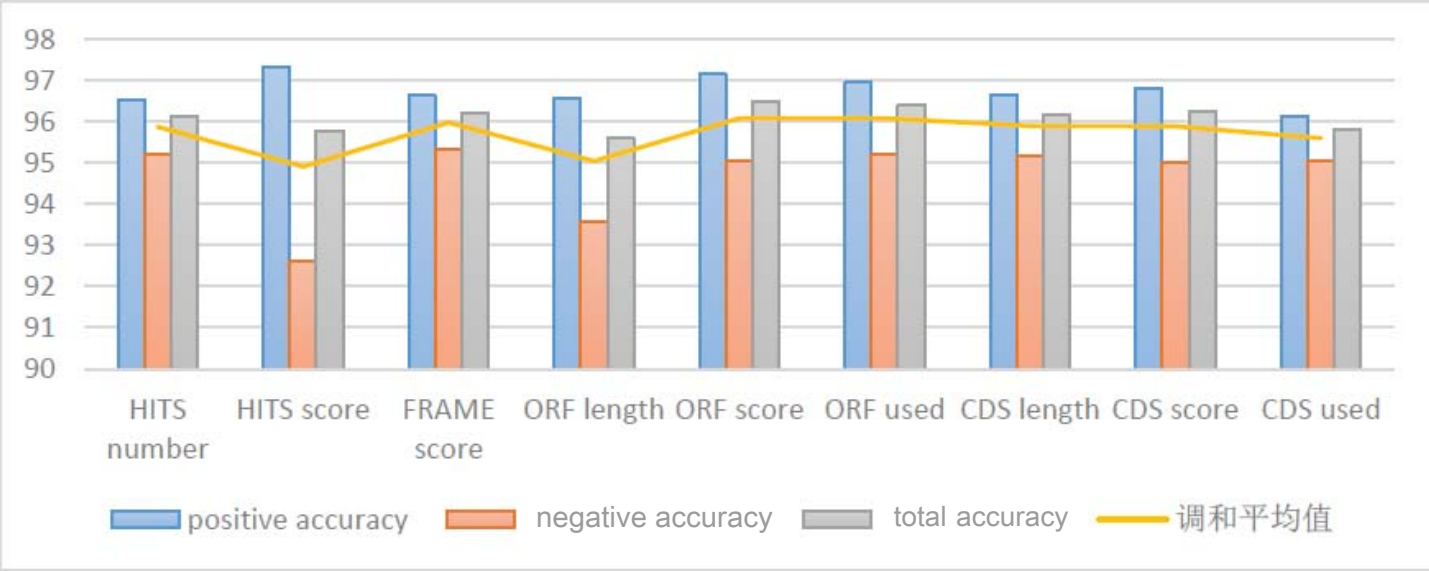


SBS

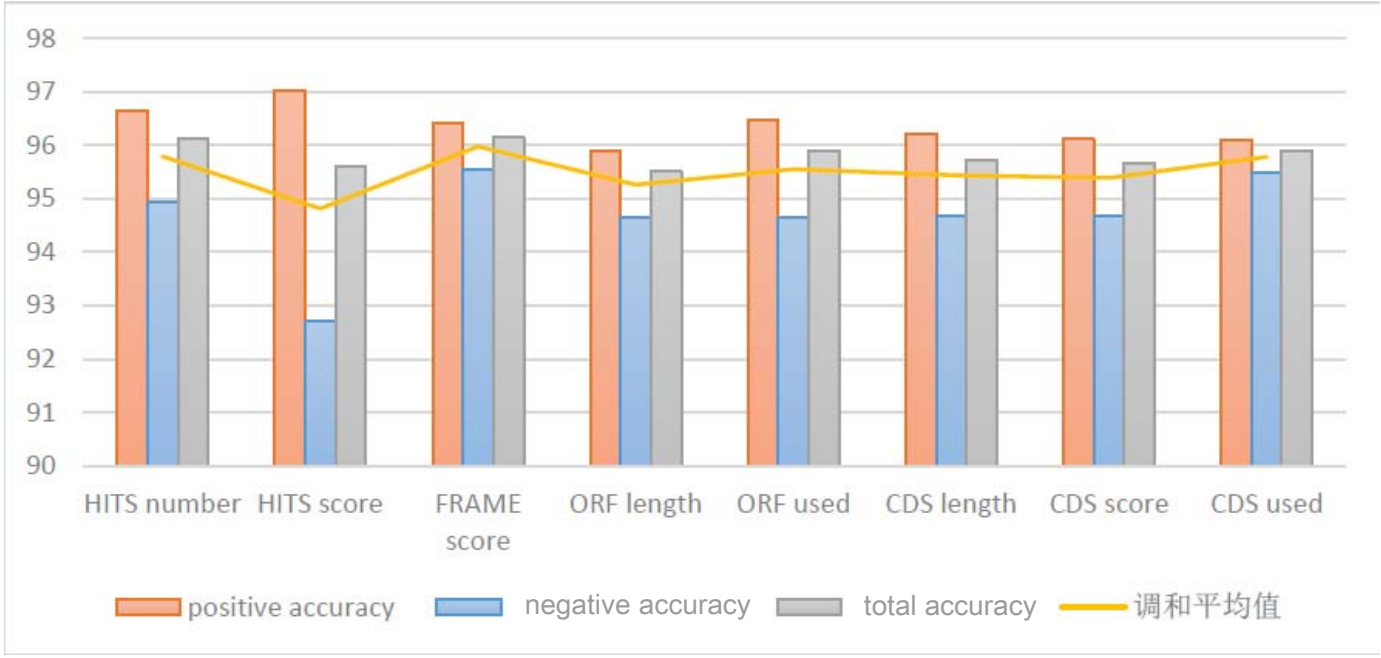
Set of variables starts from an set which has all variables ,each time we remove a variable from the subset and the optimal solution in the evaluation is selected.



	all	delete OF	提高（%）
positive accurac	96.77419	96.80099	0.026796
negetive accurra	94.63674	95.2545	0.617753
totle accurace	96.09479	96.31242	0.217628
调和平均值	95.36019	96.02152	0.661323



	* all	delete ORF	提高（%）
positive accu	96.80099	97.1580817	0.357092
negetive accu	95.254497	95.0397577	-0.21474
totle accurac	96.312417	96.481683	0.169266
调和平均值	96.021517	96.087246	0.065729



	* * all	FRAME sco	提高（%）
positive accu	97.15808171	96.414763	-0.743319
negetive accu	95.03975767	95.544363	0.50460498
totle accurace	96.48168299	96.143151	-0.3385322
调和平均值	96.08724605	95.977589	-0.1096567

What is clustering

- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

--from wikipedia

Distance

- Manhattan distance $d_{ij}(1) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$
- Euclidean distance $d_{ij}(2) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$
- Minkowski distance $d_{ij}(q) = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^q\right]^{1/q}$
- Chebyshev distance $d_{ij}(\infty) = \max_{1 \le k \le p} |x_{ik} - x_{jk}|$
- Mahalanobis distance $d_{ij}(M) = \sqrt{(\boldsymbol{x}_{(i)} - \boldsymbol{x}_{(j)})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_{(i)} - \boldsymbol{x}_{(j)})}$
- Lance and Williams distance $d_{ij}(L) = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}$

Change to distance

- Using R
- **dist(x, method =“euclidean”, diag = FALSE, upper = FALSE, p=2)**
- **x** a numeric matrix, data frame or "dist" object.
- **method** the distance measure to be used. This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski". Any unambiguous substring can be given.
- **diag** logical value indicating whether the diagonal of the distance matrix should be printed by print.dist.
- **upper** logical value indicating whether the upper triangle of the distance matrix should be printed by print.dist.
- **p** The power of the Minkowski distance.

Hierarchical clustering method

- Single linkage method $D_{\mathrm{MJ}} = \min\{D_{\mathrm{K}J}, D_{L\mathrm{J}}\}$
- Complete linkage method $D_{\mathrm{MJ}} = \max\{D_{\mathrm{K}J}, D_{L\mathrm{J}}\}$
- Median method ${D_{MJ}}^2 = \frac{1-\beta}{2}\left({D_{KJ}}^2 + {D_{LJ}}^2\right) + \beta {D_{KL}}^2$
- Average linkage method $D_{\mathrm{MJ}} = \frac{n_K}{n_M} D_{K\mathrm{J}} + \frac{n_L}{n_M} D_{L\mathrm{J}}$
- Centroid method ${D_{MJ}}^2 = \frac{n_K}{n_M} {D_{KJ}}^2 + \frac{n_L}{n_M} {D_{LJ}}^2 - \frac{n_K n_L}{{n_K}^2} {D_K}^2$
- Ward method ${D_{MJ}}^2 = \frac{n_J+n_K}{n_J+n_M} {D_{KJ}}^2 + \frac{n_J+n_L}{n_J+n_M} {D_{LJ}}^2 - \frac{n_J}{n_J+n_M} {D_{KJ}}^2$

hclust

- hclust(d, method = "complete", members = NULL)
- d a dissimilarity structure as produced by dist.
- method the agglomeration method to be used. This should be (an unambiguous abbreviation of) one of "ward", "single", "complete", "average", "mcquitty", "median" or "centroid".

Reference

- 统计建模与R软件
- http://www.cnblogs.com/xiangshancuizhu/archive/2012/03/12/2392360.html
- http://en.wikipedia.org/wiki/Feature_selection
- http://en.wikipedia.org/wiki/Cluster_analysis
- http://www.biostars.org/p/14156/

Thank you for your attention