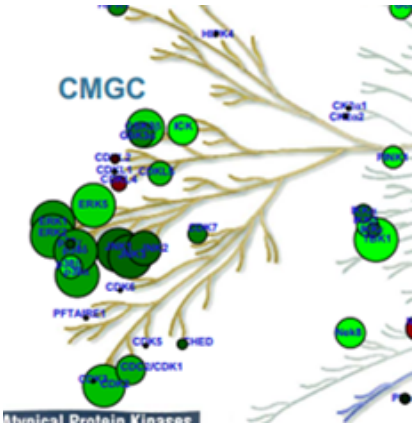
Background to the Upstream Kinase Analysis PamApp

Upstream Kinase Analysis using phosphoNET



This page is background information to the following PamApps:

- PamApp for PTK Upstream Kinase Analysis
- PamApp for STK Upstream Kinase Analysis

In the New 2018 upstream kinase PamApp How to use the 2018 Upstream Kinase PamApp, there are additional Databases that are used to derive the putative Kinase-peptide relationships. These databases are given higher weight, and higher Rank of 0. The principle of the statistics and calculations used is as described below

Also see:

- How to use the Upstream Kinase Analysis PamApps.
- BioSB 2015 Upstream Kinase Analysis poster

For other methods for the interpretation of kinomic profiles, see: Biological analysis.

Table of Contents

Introduction

Kinase phosphorylation activities measured on peptides (which show different responses) can be further evaluated by relating them to kinases that are able to phosphorylate a particular site on the peptide sequences, typically according to PTM databases or PTM prediction tools. In this way information may be obtained on the *upstream kinases* that are responsible for relevant (changes in) phosphorylation activity in an experiment.

The PamApp for PTK Upstream Kinase Analysis and the PamApp for STK Upstream Kinase Analysis implement a systematic and integrated way to analyze the multiple parallel changes in peptide phosphorylation observed in experiments in which two groups are compared, i.e. the type of experiments for which the standard statistical analysis is implemented in the PamApp for Two Group Comparison. The upstream kinase analysis provides information on which kinases are responsible for the changes between the groups.

The analysis is based on documented kinase-substrate relationships (from *in vivo / in vitro* experiments) complemented with ''in silico'' predictions for the upstream kinases of phosphorylation sites in the human proteome that are retrieved from the phosphoNET database. For the latter, a prediction algorithm was derived from known interactions between kinases and phosphorylation site. The prediction algorithm was then used to predict the strength of undocumented interactions.

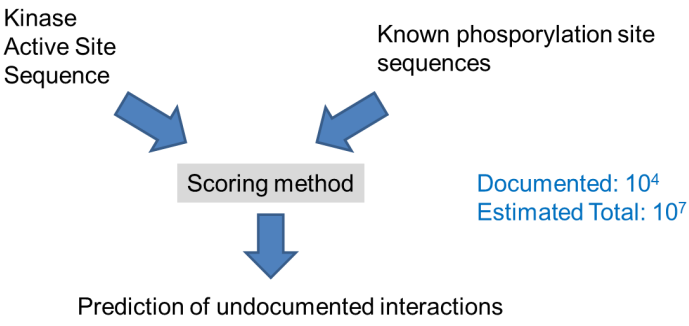


Figure 1

For this analysis, we have use a comprehensive *in silico* prediction database rather than literature databases only because it is clear that only a minor part of the kinase-phosphosite interactions have been documented in literature. Therefore, limiting to literature data only would be too stringent and result in a bias towards well studied kinases and phosphosites. Obviously, this extrapolation of kinase-phosphosite interactions comes at the expense of the confidence that can be placed in the results. The method is therefore complementary to more literature based methods like pathway analysis: more explorative but less limited by issues like limited coverage and bias that are common to literature databases.

We use the phosphoNET data base to map the PamChip peptides to putative upstream kinases. Then, an analysis of the difference between experimental groups is performed in terms of the upstream kinases rather than in terms of peptides. This *integration* of prior information and experimental data is brought about by grouping the peptides in non-exclusive sets that "belong" to individual kinases. Based on the collective change between these peptide set the corresponding upstream kinases are scored for potential involvement in the difference between groups.

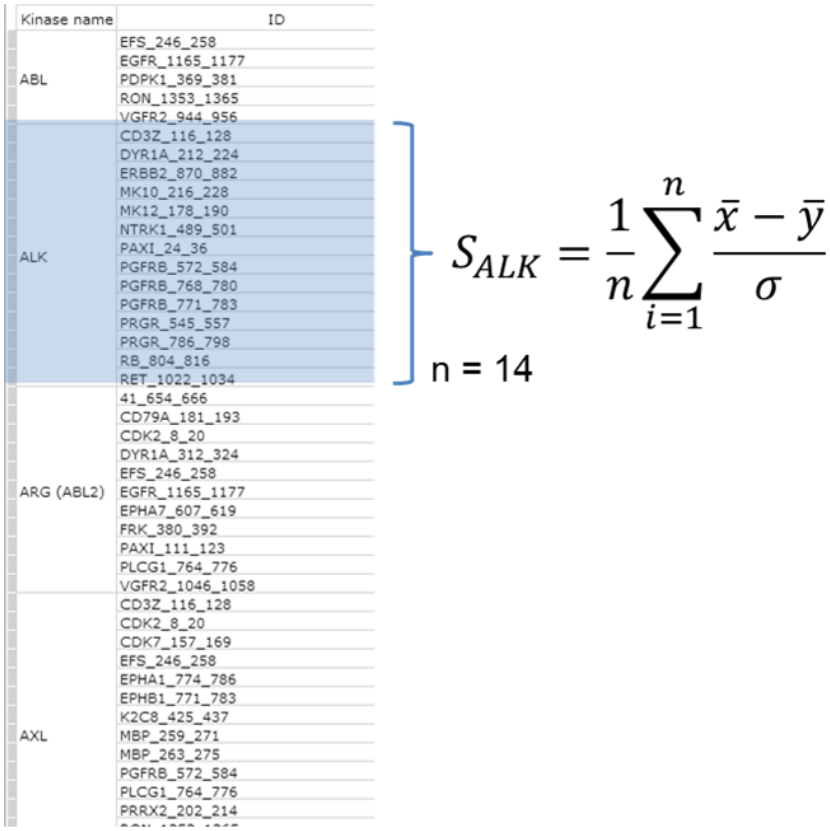


Figure 2

Figure 2 depicts a possible construction of peptide groups for some example kinases. A *(normalized) kinase statistic* $\boldsymbol{S_k}$ for the change in phosphorylation between samples in group x and group y can be calculated as the mean signal to noise ratio (SNR) of the individual peptides in the group: the mean signal in x minus the mean signal in y divided by the standard deviation. Hence, in this example the value of S_k for ALK would be the mean SNR of the 14 peptides associated with this kinase. If there is an, on average, larger change of the peptides in the same direction (i.e. all upwards or all downwards), a larger absolute value of S_k would result.

- The *significance* of S_k can be determined by a permutation test, randomly permuting the samples between the groups. Hence, the *significance* assesses if the absolute value of S_k using the experimental groups is higher than what can be expected using a random grouping of the samples.

- The *specificity* is determined by a permutation test, randomly perumting the peptides between the peptide sets. Hence, the *specificty* assesses if the value of S_k is associated with a specific set of peptides, or that similarly high values are likely to be obtained using random sets of peptides.

Based on scoring *significance* and *specificity* of S_k in this way, kinases may be ranked according to their putative involvement in the observed experimental differences.

Methods

Mapping peptides to kinases

Peptides on PamChip represent phospho-sites for tyrosine or serine-threonine kinases. Using standard BLAST tools the peptides on the chip were matched to pospho-sites in the human proteome. The phosphoNET database was then queried for kinases that are predicted to phosphorylate these sites.
When experimental data comparing two experimental groups is input in a PamApp for upstream kinase analysis, this information is added to the included peptides by enrichment.

A **peptide** is enriched with an **upstream kinase** if

- the peptide has a match to a phospho-site for which the upstream kinase is found in an included *in vivo / in vitro* database.
- the peptide has a match to a phospho-site for which the upstream kinase has a phosphoNET version 2 score larger than 300.

For each peptide this results in a list of kinases, ranked according to the phosphoNET score. Peptides found in an *in vivo / invitor* database will always get the highest rank (0). Peptides for which no phosphosite-upstream kinase link is found are silently removed from the data set.

Input data

- *n* x *k* data matrix **X** containing data for *n* observations (samples) of *k* peptides. **X** typicaly contains log-transformed kinase activity profiles, or log fold changes if compound effect profiles are analyzed.
- *n* x 1 grouping vector **y** , containing the binary grouping for each observation. $y_i = 0$ if observation *i* belongs to group 1, $y_i = 1$ if observation *i* belongs to group 2.
- The *Upstream Kinase Basis*, a *k* x *u* matrix **C** containing the mapping of *k* peptides to *u* upstream kinases. $C_{ij} = 0$ unless for peptide *i* kinase *j* appears as an upstream kinase, and kinase *j* has rank <= N_k for peptide *i*, where N_k is an integer, typically between 5 and 15 (see below). If C_{ij} is non-zero its value can be 1 (to signifiy the presence of kinase *j* for peptide *i*), or it can contain some weighting value, i.e. the phosphoNET score.

Per peptide difference statistic

For each peptide a difference statistic *p* is calculated, the following options have been implemented:

Name	Description
Signal to noise ratio (SNR)	$p = \frac{\bar{x}_{y=1} - \bar{x}_{y=0}}{\sqrt{\sigma^2_{y=1} + \sigma^2_{y=0}}}$
Identity	used when the analysis can not be formulated as a two group analysis.

with

Symbol	Description
$\bar{x}_{y=i}$	mean of the observations in group y == i
$\sigma_{y=i}$	standard deviation of observations in group y = i

Normalized kinase statistic

A *u* x 1 vector **s** containing the (normalized) kinase statistic for kinase 1..m is calculated by:

$\mathbf{s} = \mathbf{b} * (\mathbf{C}^T.\mathbf{p})$

where:

- **p** is a k x 1 vector containing the per peptide difference statistic for peptide 1..k.
- **b** is a u x 1 normalization vector with $b_i = 1/a_i$, *i* = 1..u, with a_i the number of non-zero elements in column i of **C**.
- Superscript ^{T} indicates matrix transpose, * indicates element by element multiplication.

i.e. **s** contains an average of the per peptide difference statistic **p**, in which the latter is weighted by **C**.

Significance Score

The significance of the multiple kinase associated statistic values in **s** is assessed by a permutation test invovlving permutations of the observations. A "permuted version" of **s**, $\mathbf{s_p}$ is calculated *M* times by repeating the above procedure for calculating **s** with the group membership of the samples randomly reassigned. Subsequently, the Significance Score Q_{sg} for each element *s* of **s** is assessed by counting the number of times *m* that $|s_p| >> |s|$:

$Q_{sg} = -^{10} \log\{ \max(m/M, 1/M) \}$

- If the number of observations *n* < 7, all possible permutations are generated with maximum of *M* = 500 (randomly selected) permutations.
- Otherwise, a set of *M* = 500 random permutations is used.

Specificity Score

The specificity score Q_{sp} is calculated in the same way as Q_{sg} , but now on the basis of *M* = 500 random permutations of the peptides.

Final score

The final score Q is used to finally rank the included kinases for putative involvement in the obserevd experimental differences. Typically, for larger n

$Q = Q_{sp} + Q_{sg}$

For smaller n the permutations for Q_{sp} become too few and:

$Q = Q_{sp}$

Where a higher Q means a higher ranking in the putative upstream kinase list. Usually it is observed the effect on the final ranking of changing between both ways of calculating Q is minor, as the ranking is mainly determined by Q_{sp} in the first place. Typically, a relatively large number of kinases has a large Q_{sg} , of which only a subset depends on its specific set of peptides.

What value of N_k to use?

The main free parameter in that can influence the results of the analysis is N_k , the maximum rank (or number) of upstream kinases to be "added" to a peptide. Therefore, the analysis is performed for a range of values of N_k , typically N_k = 4,5,...,12. The final result is then reported for all these analyses simultaneously. This can be used to assess if the ranking of kinases depends strongly on the choice of N_k . Based on this a researcher can e.g. deprioritize kinases when a high ranking depends strongly on a particular choice of N_k . See How to use the Upstream Kinase Analysis PamApps for details on the analysis reports.

Bionavigator Implementation

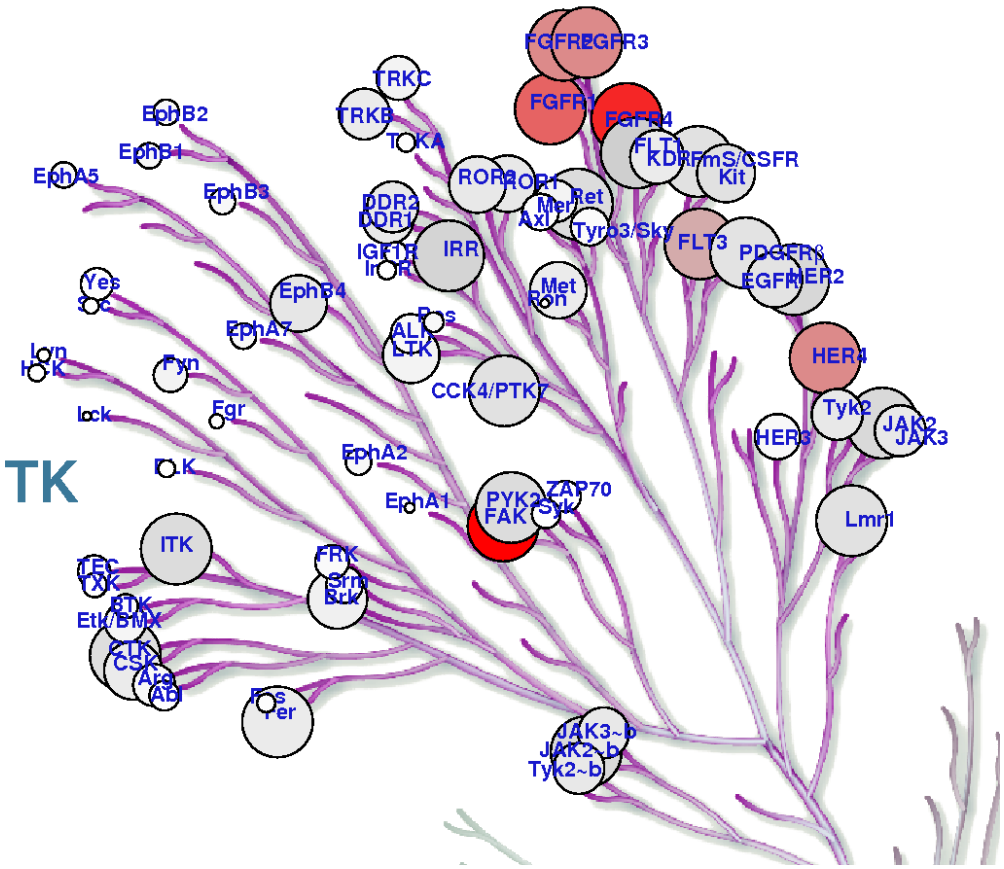
The analysis has been implemented in R (3.3.1), and interfaced to Bionavigator. For general use the analysis is packaged as a Step PamApp:

- PamApp for PTK Upstream Kinase Analysis
- PamApp for STK Upstream Kinase Analysis

Results

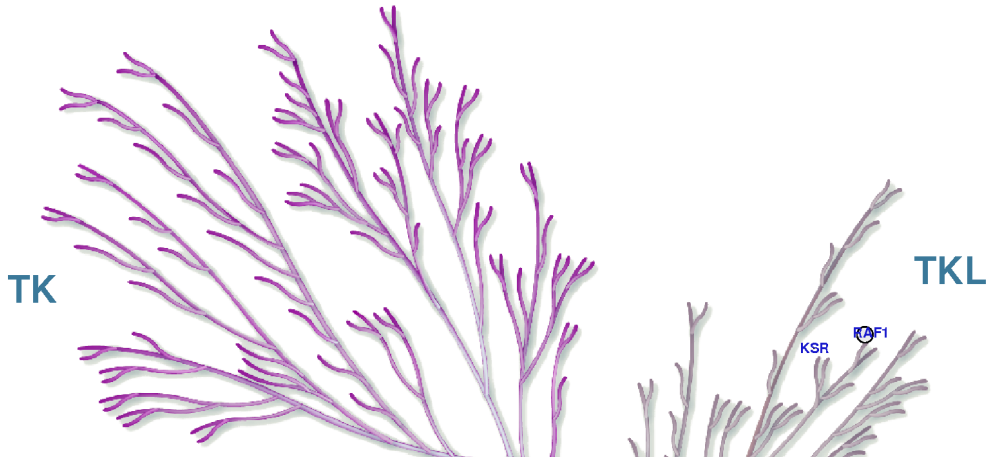
FGFR example

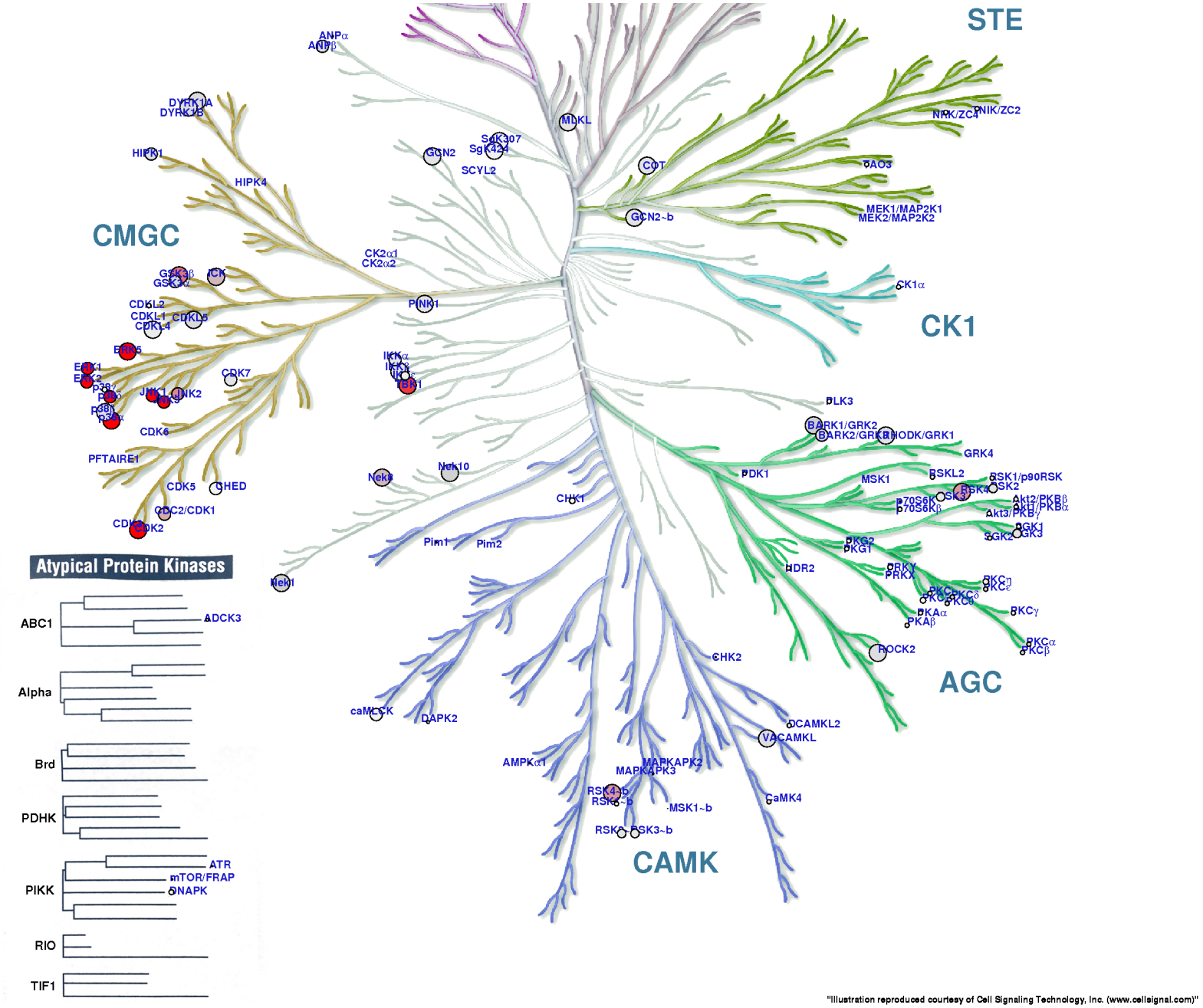
The figure below shows the Upstream Kinase Analysis example of comparing FGFR2 over-expressing cell lines to non over-expressing cell lines. Results are mapped to the kinome phylogenetic tree, of which only the PTK part is shown. Size of the circles indicate the significance score, color indicates the specificity score (red = more specific). "Clustering" of high scores on and near the FGFR branch can be observed.



MAPK example

As an STK example, the figure below shows the mapping of Upstream Kinase Analysis resulting from comparing cell lines with / without activation of MAPK/Erk signaling downstream of RAF-1.





Supplementary

Presentation on Upstream Kinase Analysis

- Upstream Kinase Analysis app.pptx
- BioSB 2015 Upstream Kinase Analysis poster

phosphoNET phosphorylation predictor

- KinasePred-BIBM-10.pdf
- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3379035/

More on functional class scoring

- http://www.pnas.org/content/102/38/13544.full.pdf+html
- http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002375

Similar implementation at Max Planck Institute for Molecular Genetics

- http://phoxtrack.molgen.mpg.de/

From a Report (RH)

The kinase phosphorylation activities measured on peptides (which show different responses) can be further evaluated by relating them to kinases that are able to phosphorylate a particular site on the peptide sequences. PhosphoNet (www.phosphonet.com) provides an algorithm to predict the phosphorylation site specificity of 478 human protein kinases based on the primary structure of the catalytic domain of these kinases. For many kinases, this has been validated with known kinases and their substrates, whereas for poorly investigated kinases the preferred substrate sequences are deduced from the primary structure by the algorithm. A data analysis module was developed that uses this information to determine for each peptide which kinase are likely to phosphorylate a particular site. For each peptide that differentiates two groups, such an analysis is performed. The putative kinases differentiating the groups are ranked based on the specificity (compare the observed difference with the differences obtained when the peptide labels are permutated) and the sensitivity (compare the observed difference with the difference obtained when the sample labels are permutated). The outcome is a list of putative upstream kinases. Several variables in the analysis algorithm determine the outcome. One of these is the number of putative upstream kinases that are included in the analysis, for many sites are phosphorylated by many kinases, albeit with differing efficiencies. Another is the weight factor for ranking the upstream kinases. Since families of kinases have similar substrate preferences, several members of a family may appear at a high position in the list. Further experiments are required to confirm the involvements of the kinase(s) indicated, for a high ranking kinase might not be present in the samples under investigation.

phylogenetic tree web service

External Web services for mapping annnotation (such as upstream kinase analysis scores) to the Manning et al. (Science 298 1912-34) phylogenetic tree of the kinome.

- Kinome render (Obsolete, not available anymore)
- KinMap
- Coral

An annotation file that can be used to create such images with the Kinome Render web service is automatically created by the Upstream Kinase Analysis tools.

Also see: PamApp for Kinome Renderer File Parser

UpstreamKinaseTag