Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs

John C. Obenauer, Lewis C. Cantley and Michael B. Yaffe*

Center for Cancer Research, E18-580, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA and 
Division of Signal Transduction, Beth Israel Deaconess Medical Center, Boston, MA, USA

Received February 12, 2003; Revised and Accepted April 7, 2003

ABSTRACT

Scansite identifies short protein sequence motifs that are recognized by modular signaling domains, phosphorylated by protein Ser/Thr- or Tyr-kinases or mediate specific interactions with protein or phospholipid ligands. Each sequence motif is represented as a position-specific scoring matrix (PSSM) based on results from oriented peptide library and phage display experiments. Predicted domain-motif interactions from Scansite can be sequentially combined, allowing segments of biological pathways to be constructed in silico. The current release of Scansite, version 2.0, includes 62 motifs characterizing the binding and/or substrate specificities of many families of Ser/Thr- or Tyr-kinases, SH2, SH3, PDZ, 14-3-3 and PTB domains, together with signature motifs for PtdIns(3,4,5)P3-specific PH domains. Scansite 2.0 contains significant improvements to its original interface, including a number of new generalized user features and significantly enhanced performance. Searches of all SWISS-PROT, TrEMBL, Genpept and Ensembl protein database entries are now possible with run times reduced by ~60% when compared with Scansite version 1.0. Scansite 2.0 allows restricted searching of species-specific proteins, as well as isoelectric point and molecular weight sorting to facilitate comparison of predictions with results from two-dimensional gel electrophoresis experiments. Support for user-defined motifs has been increased, allowing easier input of user-defined matrices and permitting user-defined motifs to be combined with pre-compiled Scansite motifs for dual motif searching. In addition, a new series of Sequence Match programs for non-quantitative user-defined motifs has been implemented. Scansite is available via the World Wide Web at http://scansite.mit.edu.

INTRODUCTION

Characterizing protein interactions on a proteome-wide scale is required to catalyze the advance of systems biology. Online databases of protein sequences (1–5) and known protein–protein interactions (6–8) are the first steps taken in this direction, but finding new interactions will require new combinations of experimental and computational methods. Scansite (http://scansite.mit.edu) is a computational tool built on experimental binding and/or substrate information from oriented peptide library screening (9–13) and phage display experiments (14), together with detailed biochemical characterization to derive a weight matrix-based scoring algorithm that predicts protein–protein interactions and sites of phosphorylation (15).

DOMAINS AND MOTIFS

The accumulated molecular structures in the Protein Data Bank (PDB) make it clear that eukaryotic proteins are often built with a modular architecture, combining domains that fold and function independently into larger polypeptides. These domains often occur in multiple unrelated proteins, where they fulfill similar targeting functions. Identification of these domains within a protein can be a valuable indicator of the function of the protein as a whole and can assist in placing that protein within the correct cell signaling pathway. A number of modular domains such as WW, SH2, SH3, PTB, PDZ and 14-3-3 bind to their ligands through direct interactions with very short amino acid sequences (typically <10 amino acids), or in the case of protein kinases, phosphorylate a Ser-, Thr- or Tyr-containing sequence motif in their protein substrates. Modular binding domains are typically fairly long (60–300 residues) and can be identified using sequence comparison methods and Hidden Markov Models [e.g. Pfam (16) and SMART (17)]. In contrast, the corresponding motifs to which they bind are much shorter (3–10 residues) and have been more elusive to locate. The current release (version 7.8) of Pfam, for example, identifies 4941 protein domains and families, but only 18 motifs (16). Scansite was developed to address this need and to facilitate work in our own laboratories on signaling by protein...
kinases and modular phosphopeptide- and phospholipid-binding domains.

Many of the motifs in Scansite were determined using oriented peptide library experiments. In this technique, degenerate peptides with a single fixed (orienting) central residue are incubated with one type of domain (9–13). Because of our laboratories’ research focus, this central residue was typically a Ser, Thr or Tyr for protein kinase domains, or a phosphoSer/Thr or phosphoTyr residue for phosphospecific binding domains (such as SH2, PTB or 14-3-3 domains). Peptides that were phosphorylated by the kinase or were bound by the binding domain were isolated and sequenced as an ensemble by Edman degradation. When sequenced in this manner, each Edman cycle reveals the relative amount of each amino acid residue occurring at that position. This information is then scaled and normalized to produce a scoring matrix (i.e. a PSSM) which quantitatively indicates the preference for each amino acid type at each position within the domain’s recognition motif. These matrices can then be used to score entire databases of protein sequences to find a small number of proteins with high-ranking motif matches, indicating possible protein–protein interactions. As the number of motifs grew, the opposite search became practical as well: scanning a single protein sequence for matches to any of the motifs in our database.

We have collected these programs to create a user-friendly web-based tool accessible to the entire scientific community that allows investigators to search for motifs recognized by commonly occurring domains within a protein sequence query of their choice or to search entire protein sequence databases for optimal motif matches. The Motif Scan ensemble of programs computationally identifies all motifs within a given user-specified protein, while the Database Search ensemble of programs finds all proteins in a protein database, such as SWISS-PROT, that match a given motif. By repeated queries using the results of one search to launch another, it is possible to infer several steps of a signaling pathway in silico. For example, if a newly discovered protein is predicted by Scansite to be phosphorylated by the kinase domain from Akt and the resulting phosphorylation is predicted to create a binding site for 14-3-3 proteins, then the newly discovered protein is likely to function in a signaling pathway involving these proteins. These types of analyses performed on protein sequence databases can functionally annotate a limited number of promising interactions that merit experimental investigation and may also suggest that other intermolecular interactions are unlikely, at least within the limits of sequence-based prediction.

**STRINGENCY LEVELS**

Threshold values need to be assigned when scanning query proteins with the Motif Scan programs to decide which scores are likely to suggest real interactions. Scansite incorporates three settings, labeled ‘high’, ‘medium’ and ‘low’ stringencies; the high stringency setting is the most restrictive and reports a ‘hit’ only if the score falls within the top 0.2% of scores when the motif matrix of interest was applied to the vertebrate subset of SWISS-PROT. This dataset was chosen as a reference because of the non-redundant nature of SWISS-PROT and the relevance of vertebrate proteins to the type of cell signaling events predicted by Scansite. These values were found to increase the reliability of prediction of true positive ‘hits’ while minimizing the number of predicted false negative interactions, based on a comparative analysis of mammalian and bacterial database subsets (15). The medium and low stringencies were then arbitrarily chosen at 1 and 5%, respectively.

Scoring percentiles in the Database Search programs, on the other hand, are calculated de novo, based solely on the protein database subset selected for the search. For example, a search among human proteins will yield sites whose percentiles are relative to all human proteins included in the search. The same site can thus have a different percentile for different database searches, but its score is always constant. It should always be borne in mind by the user that Scansite predictions are based solely on 1D sequence comparison and all predicted interactions must be experimentally verified before they can be considered valid.

**MATERIALS AND METHODS**

**Server**

The public collection of Scansite programs runs on a Dell PowerEdge 8450 server, with 8 Intel Xeon 733 MHz CPUs and 4 Gb of RAM. Two 32 Gb hard drives are used in a RAID 1 array. The operating system is Red Hat Linux 7.3.

**Development**

All development for Scansite version 2.0 was performed using the GNU GCC compiler, the PHP 4.0 and Perl 5.5 interpreters, Mandake Linux 8.0 through 9.0, Red Hat Linux 7.1 through 7.3, the Apache 1.3 web server, the MySQL 3.23 relational database and the KDE desktop environment.

**SCANSITE PROGRAMS**

A total of 10 programs are included in Scansite 2.0 and these are listed in Table 1. The Motif Scan programs can accept either a protein accession number or a sequence as input and can optionally accept a user-defined motif. The Database Search programs can operate on one or more Scansite motifs, one or more user-defined motifs or combinations of Scansite and user-defined motifs. The Quick Matrix Method allows users to construct a roughly quantitative matrix based on qualitative residue preferences for a sequence motif. The Sequence Match programs allow users to find occurrences of one or two specified consensus sequences in the protein databases and can also be used to find any MySQL-recognized regular expression. A brief description of using each of these programs follows. More detailed instructions can be found in the tutorial on our web site (http://scansite.mit.edu/tutorial/tutorial.html) (see also 18).

**Motif Scan**

To use the Motif Scan programs, users should go to the web site http://scansite.mit.edu. Under the heading ‘Motif Scan’,
Table 1. List of programs available on the Scansite web site

<table>
<thead>
<tr>
<th>Motif Scan</th>
<th>Search a protein sequence for motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scan a Protein by Accession Number or ID</td>
<td>Takes accession number or ID as input (e.g. RB_HUMAN, P06400)</td>
</tr>
<tr>
<td>Scan a Protein by Input Sequence</td>
<td>Takes sequence as input</td>
</tr>
<tr>
<td>Scan Input Sequence with an Input Motif</td>
<td>Takes sequence and an user-defined motif as input</td>
</tr>
<tr>
<td>Database Search</td>
<td>Searches a database for motifs</td>
</tr>
<tr>
<td>Search Using a Scansite Motif</td>
<td>Searches for a single pre-compiled Scansite motif</td>
</tr>
<tr>
<td>Search Using an Input Motif</td>
<td>Takes a user-defined motif as input</td>
</tr>
<tr>
<td>Search Using Quick Matrix Method for Making a Motif</td>
<td>Takes a semi-quantitative user-defined motif as input</td>
</tr>
<tr>
<td>Search Using Multiple Motifs</td>
<td>Searches for multiple pre-compiled or user-defined motifs</td>
</tr>
<tr>
<td>Sequence Match</td>
<td>Retrieves all sequences matching an input pattern exactly</td>
</tr>
<tr>
<td>Search Databases for Sequence Pattern</td>
<td>Takes a single sequence pattern as input</td>
</tr>
<tr>
<td>Search Databases for Two Sequence Patterns</td>
<td>Takes two sequence patterns as input</td>
</tr>
<tr>
<td>Search Databases for Regular Expression</td>
<td>Takes a regular expression as input</td>
</tr>
</tbody>
</table>

click ‘Scan a Protein by Accession Number or ID’ to use a protein from a public database or click ‘Scan a Protein by Input Sequence’ to enter a protein sequence directly. The required inputs are then displayed, which include the protein’s accession number and database of origin (or with the input sequence version, the protein’s sequence and an arbitrary name for it), followed by the list of motifs to scan for. The default setting is to search for occurrences of all motifs in the Scansite database. Alternatively, one or more individual motifs can be selected, or several motifs of similar type (i.e. a ‘motif group’) can be selected at once. The list of motifs currently available in Scansite is shown in Table 2. Users can search at high stringency (the default choice), which shows only the strongest motif matches or at medium or low stringency to see weaker sites. Finally, users can elect to identify domains in the protein sequence, which Scansite accomplishes by parsing the results from an external call to the Pfam server at Washington University, St Louis (16). This lengthens the time needed to generate results, but the domain information is often very informative. With all these settings selected, clicking the ‘Submit Request’ button initiates the scan. The result will show a schematic map of the protein with the predicted sites found (Fig. 1) and a detailed table showing the score and sequence of each one (Fig. 2).

**Database Search**

To use the Database Search program, users should click ‘Search Using a Scansite Motif’ under the ‘Database Search’ heading. A list of all the motifs in Scansite is shown. Users should select one of the motifs to search with and select the name of the protein database to search. The databases currently available are SWISS-PROT, TrEMBL, Genpept and Ensembl. Optionally, the search can be limited to proteins in just one species or a category of organisms, including mammals, vertebrates, invertebrates, plants, fungi, viruses and bacteria and archaea (grouped together). Other options allow searching within a specified range of molecular weights and isoelectric points, to facilitate comparison with two-dimensional gel electrophoresis experiments. Restricting the results by keywords in the protein description and/or by characteristic subsequences is also possible. The last user-specified parameter is the desired size of the search output, ranging from 50 to 2000 reported sites. Clicking ‘Submit Request’ starts the search. The resulting table (Fig. 3) lists all sites found, identifying the associated protein’s name, description, sequence, molecular weight and isoelectric point. Any protein found from a database search can be rapidly submitted to the Motif Scan program by clicking the ‘Submit’ button on the far left of each output line.

In addition to the pre-compiled Scansite motifs listed, investigators can use their own motifs to search databases, using the program ‘Search Using an Input Motif’. A tab-delimited text file containing a weight matrix is uploaded into Scansite and the subsequent options and output are the same as described above. Instructions on how to create and upload a matrix are provided in the tutorial page on our website (http://scansite.mit.edu/tutorials/tutorial.html).

One variation on the Database Search is the program ‘Search Using Quick Matrix Method’. This program allows users to define an approximate motif by specifying a short pattern of amino acids, where wildcards are allowed. For a motif such as RXSXL, this sequence can be entered in the row of positions labeled ‘Primary Preference’. Optionally, if it was known that proline can substitute for the leucine, a P can be entered in the ‘Secondary Preference’ row at that position. Scansite makes a crude weight matrix based on these inputs, assigning a score of 9.0 to residues in the primary preference row, a score of 4.5 to those in the secondary preference row and a score of 1.0 to all unspecified residues. The results of using the Quick Matrix Method will be less quantitative than a normal database search, but can yield useful results when only limited motif information is available.

**Sequence Match**

The Sequence Match programs are new in the current release of Scansite. As with the Quick Matrix Method, these programs are useful when only partial motif information is available. Unlike the Quick Matrix Method, these programs do not provide quantitative match ranking, but they instead retrieve all proteins in a database that exactly match the sequence pattern specified, similar to the programs Patscan (Ross Overbeek and Alex Rodriguez, http://www-unix.mcs.anl.gov/compbio/PatScan/HTML/patscan.html) and ScanProsite (http://us.expasy.org/tools/scanprosite/). Unlike those two programs, Sequence Match will accept the widely used regular expression syntax common in Perl, PHP, MySQL and other programming languages. By choosing a motif, users can instruct Scansite to scan a protein sequence for motifs.
environments. This kind of information can help an investigator decide how rare or specific a hypothetical motif is, how functionally similar the proteins containing the motif are, whether a motif occurs more commonly in one species or another and how many proteins may cross-react with an antibody made using the motif as an epitope. As with the Database Search, the proteins retrieved can be limited to the most relevant ones by specifying a single species, molecular weight range and values for the other options mentioned previously.

There are three Sequence Match programs. The first and simplest takes a single consensus sequence as input, which may contain wildcards. The second program looks for two different consensus sequences occurring simultaneously in the same protein. The third and most flexible program is ‘Search Databases for Regular Expression’. Unlike the first two programs, this program allows gaps of any length, alternative residues at any position and motifs at the N- or C-termini of proteins (such as signal sequences or antibody epitopes). Any regular expression recognized by MySQL can be used as the search term and our web site gives the full list of allowed symbols as well as several biologically useful examples.

**Table 2. Current list of motifs included in Scansite**

<table>
<thead>
<tr>
<th>Motif Type</th>
<th>Motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phosphoserine/threonine binding domains</td>
<td>14-3-3 mode 1</td>
</tr>
<tr>
<td>Tyrosine kinase domains</td>
<td>Abl, EGFR, FGFR, Insulin receptor, Itk, Lck, PDGFR, Src</td>
</tr>
<tr>
<td>Src homology 2 domains</td>
<td>Abl, Crk, FGFR, Fyn, Grb2, Itk, Lck, Nck, p85, PLC γ (C terminal SH2), PLC γ (N terminal SH2), Src, SHIP</td>
</tr>
<tr>
<td>Src homology 3 domains</td>
<td>Abl, Amphinysin, Cbl-associated protein, Cortactin, Crk, Grb2, Intersectin, Itk, Nck, p85 α (mode 1), p85 α (mode 2), PLC γ, Src</td>
</tr>
<tr>
<td>Basophilic serine/threonine kinase domains</td>
<td>Akt, Calmodulin-dependent kinase 2, Clk2, Protein kinase A, PKC α/β/γ, PKC δ, PKC ε, PKC μ, PKC γ</td>
</tr>
<tr>
<td>DNA damage kinase domains</td>
<td>ATM, DNA protein kinase</td>
</tr>
<tr>
<td>Acidophilic serine/threonine kinase domains</td>
<td>Casein kinase 1, Casein kinase 2, GSK3</td>
</tr>
<tr>
<td>Proline-dependent serine/threonine kinase domains</td>
<td>Cdc2, Cdk5, Erk1, p38 MAP kinase</td>
</tr>
<tr>
<td>Kinase binding domains</td>
<td>Erk1, PDK1</td>
</tr>
<tr>
<td>PDZ binding domains</td>
<td>PDZ class 1, PDZ class 2, PDZ (nNOS) class 1, PDZ (nNOS) class 3</td>
</tr>
<tr>
<td>Phosphotyrosine binding domains</td>
<td>Src</td>
</tr>
<tr>
<td>Lipid binding domains</td>
<td>PIP3-binding PH</td>
</tr>
</tbody>
</table>

**Figure 1.** Description of elements in Motif Scan graphical output. The protein query (in this case, the transcription factor FOXO1) is represented schematically as a line, with colored rectangles marking known domains. Labels above the protein indicate where motifs were found and identify the motif family. Labels below the protein indicate the name and range of each domain found. If the protein’s annotation includes phosphorylation sites that have been experimentally mapped (generally true only for some SWISS-PROT entries), these are also indicated below the domains. On the next line, a plot of the predicted surface accessibility at each residue, calculated using a 6 amino acid running window (19) is shown. The ruler at the bottom marks numbered intervals along the protein sequence.

**IMPROVEMENTS IN VERSION 2.0 OF SCANSITE**

**Speed**

Program execution speed has been significantly improved for the Database Search programs. Storing protein sequence information in a relational database rather than in text files, in
combination with rewriting the base code, shortened the time needed for a typical database search by approximately a factor of three compared with Scansite version 1.0. Our protein sequence databases are currently updated with each major release of Genpept, SWISS-PROT, TrEMBL and Ensembl. Between updates, very recent additions to these databases may not be present in Scansite.

**Targeted searches**

In addition to speed, the MySQL relational databases for protein sequences and motif PSSMs facilitate restricted database searches based on pre-annotated database entries. Scansite 2.0 gives researchers the ability to find motifs in proteins from a single species or genus, within a range of molecular weights and isoelectric points, or containing keywords, and/or a characteristic subsequence (which can lie outside the motif region). The Motif Scan programs similarly benefit: rather than searching for all motifs or individually selected ones, users can now search by motif ‘groups’, where functionally similar motifs have been grouped together (e.g. SH2 domains, SH3 domains, tyrosine kinases and others) (Table 2). One or more motif groups can also be combined with one or more individually selected motifs.

**Graphics**

The algorithm previously used to display sites and domains graphically along the protein sequence sometimes led to overlapping text, making annotations difficult to read. The new algorithm displays many more sites and domains without overlap. In response to numerous user requests, the generated graphic is now a single downloadable PNG image to facilitate publication of users’ results.

**Two-dimensional gel electrophoresis**

Results from a Database Search can be sorted by molecular weight or isoelectric point and the search can be restricted to proteins within a narrow range of both parameters. As a result, Scansite can be used in conjunction with two-dimensional gel electrophoresis experiments to help identify spots in regions of a gel. For experiments involving primarily phosphoproteins, the expected number of phosphate groups can be specified in the Database Search options and mass and isoelectric point calculations correspondingly adjusted.

**User-entered motifs**

Users have always been able to enter their own motifs to perform Scansite searches. In version 2.0, we made three additions. First, we now allow use of matrices that lack values for one or more amino acid types by supplying default values for those positions. Second, researchers studying selenocysteine-containing proteins can now enter motifs giving a score for selenocysteine by labeling that column ‘U’, its
accepted single-letter code. Third, motifs targeting the N-terminus of a protein sequence can now be specified, using a column labeled with the arbitrarily chosen character ‘$’ (dollar sign). The ability to use C-terminal-directed motifs has existed since version 1.0 by using the ‘*’ character and is currently used in PSSMs for PDZ domains.

**Multiple motifs**

Searching for proteins that contain motifs of more than one type can be a powerful way to increase the functional relevance of database searches (15). Version 1.0 allowed users to search for two Scansite motifs or two user-entered motifs. Version 2.0 allows users to search for proteins containing up to five different motifs, which can be any combination of Scansite motifs and user-entered motifs.

**User-contributed motifs**

The **Database Search, Quick Matrix Method and Sequence Match** programs allow users to temporarily upload one or more motifs. In Scansite 2.0, we now allow researchers to submit motifs directly into the Scansite database to make them available to other users. This should contribute favorably to the number and diversity of motif types that can be searched for in protein sequence queries. However, we cannot vouch for the accuracy of user-submitted motifs. To control for this, the web site allows users the option of including or rejecting user-submitted motifs in their scans. In addition, user-submitted motifs can be individually selected along with our standard Scansite motifs when using the **Motif Scan** programs. Interested users should contact us for information on adding motifs to the Scansite database.

**Open source**

Scansite 2.0 is a completely rewritten version of the original program, developed entirely at the Massachusetts Institute of Technology. We are releasing the source code for Scansite under the terms of the GNU General Public License, version 2 (Free Software Foundation, http://www.gnu.org/licenses/gpl.txt). Researchers interested in the fine details of our score calculations and other methodologies will thus have access to them and laboratories considering writing similar web applications can use our code to get started. The PSSMs for the 62 Scansite motifs, however, remain proprietary and are not included in the released code. This policy is intended to prevent incorporation of the motifs into unauthorized commercial products. Use of the motifs on our public web site is permitted for all users, whether commercial or not. Anyone developing new features for Scansite is encouraged to submit changes back to us for inclusion in future public releases.

**FUTURE DIRECTIONS**

The revision of Scansite has produced a significantly faster and more efficient program for finding probable protein interactions. Focused searches enabled by incorporation of a relational database will help investigators target Scansite 2.0 to their own model organisms and experiments. New motifs will continue to be added to Scansite as they become available from oriented peptide library experiments. Researchers are encouraged to submit motifs of their own to our database for others to use. More specialized protein databases will be added over time, such as the RefSeq database and the mouse proteome. We are in the process of installing a second Scansite server for batch processing of long lists of sequences such as those obtained from DNA microarray experiments or genomic sequencing efforts. Future additions will include the ability to search among specific tissue types, the ability to adjust scores for predicted interactions based on their evolutionary conservation in orthologues and paralogues, the ability to restrict predicted interactions to proteins that co-localize in the same subcellular compartment, the ability to correlate predicted interactions with published data in the literature in an automated manner and the ability to automatically generate signaling network-style diagrams based on predicted interactions.

**ACKNOWLEDGEMENTS**

The authors wish to acknowledge the work done by developers who contributed to Scansite version 1.0, especially German Leparc and Stefano Volinia, as well as to members of the Yaffe and Cantley laboratories that provided the experimental data and beta-tested the programs. This work was funded by a Merck Genome Research Institute grant, the Merck/MIT Collaboration Program, NIH grants GM-60594 (M.B.Y.), GM-56203 (L.C.C.) and GM-52981 (M.B.Y.) and a Burroughs-Wellcome Career Development Award to M.B.Y.

**REFERENCES**