

nator Cycle Sequencing Ready Reaction Kit with AmpliTaq® DNA Polymerase, FS, was used under the conditions recommended by the manufacturer (PE Applied Biosystems, Foster City, Calif.). The ESTs of the present invention were generated by sequencing initiated from the 5' end of each cDNA clone.

[0259] A number of sequencing techniques are known in the art, including fluorescence-based sequencing methodologies. These methods have the detection, automation and instrumentation capability necessary for the analysis of large volumes of sequence data. Currently, the 377 DNA Sequencer (Perkin-Elmer Corp., Applied Biosystems Div., Foster City, Calif.) allows the most rapid electrophoresis and data collection. With these types of automated systems, fluorescent dye-labeled sequence reaction products are detected and data entered directly into the computer, producing a chromatogram that is subsequently viewed, stored, and analyzed using the corresponding software programs. These methods are known to those of skill in the art and have been described and reviewed (Birren et al., *Genome Analysis: Analyzing DNA*, 1, Cold Spring Harbor, N.Y., the entirety of which is herein incorporated by reference).

Example 4

[0260] Sequencing of the cDNA library LIB3373 was carried out using the primary library as the source for sequencing template. Two methods were used to isolate sequencing template: phagemid excision and single phage PCR.

[0261] In the phagemid excision method, 400-800 plaques are spread evenly over a bacterial lawn on multiple Petri plates. Blue/white selection was performed to identify putative phage containing gut inserts. White plaques were individually isolated and stored at 4° C. These are stable for several months, and thus can be isolated less frequently in greater numbers (e.g., once a month). Phagemid excision was performed in 96-sample sets from the phage stocks. This step releases the plasmid vector (containing the cDNA insert) from the Uni-Zap phage vector. This protocol was modified from Stratagene's protocol to facilitate phage adhesion and growth in 96-well culture blocks (~1.45 ml volumes). Plaques were allowed to adhere to and multiply in XL1 Blue cells co-infected with Helper phage. Cell lysis releases filamentous phage which is used to infect SOLR cells, where phagemid excision takes place. After excision, cells containing phagemids with insert were identified by a second round of selection (ampicillin resistance, blue/white colonies) immediately before isolation of DNA. Sequence-quality DNA was isolated using the Qiagen TurboPrep protocol (96-well format) and screened (EcoRI×XhoI digest) for the presence and approximate size of insert before setting up template/primer reactions. DNA sequences were then analyzed for ambiguous sequence and vector contamination and trimmed using a commercially available computer software (Sequencher), and submitted as gapped BLAST searches for comparison to public nucleotide and protein databases.

[0262] The other method utilized PCR to amplify individual inserts directly from phage; this was performed without isolation of phage DNA. The PCR reaction was carried out in 96-well format using the M13 Reverse and -20 primers. A portion of the PCR product was analyzed on an agarose gel to determine presence and size of insert. The remainder of the PCR product was purified using Qiagen's PCR Purification kit. Sequencing was then conducted using nested primers (T3/T7). This method involves a number of steps that are

analogous to the excision screening method (phage isolation, DNA purification, digest/PCR setup, agarose electrophoresis, sequence set-up) and is nearly as labor-intensive. However, this method has the potential to increase the number of clones that can be screened per week because plating is not necessary. All completed sequences were trimmed for vector contamination and ambiguous regions.

Example 5

[0263] This example illustrates sequence comparison to determine the similarity/identity of the test or query sequence with sequences in publicly available or proprietary databases. A characteristic feature of a protein or DNA sequence is that it can be compared with other known protein or DNA sequences. Sequence comparisons can be undertaken by determining the similarity of the test or query sequence with sequences in publicly available or proprietary databases ("similarity analysis") or by searching for certain motifs ("intrinsic sequence analysis") (e.g. cis elements) (Coulson, *Trends in Biotechnology*, 12: 76-80 (1994); Birren, et al., *Genome Analysis*, 1: 543-559 (1997); both of which are herein incorporated by reference in their entirety).

[0264] Similarity analysis includes database search and alignment. Examples of public databases include the DNA Database of Japan (DDBJ); Genebank; and the European Molecular Biology Laboratory Nucleotide sequence Database (EMBL).

[0265] A number of different search algorithms have been developed, one example of which are the suite of programs referred to as BLAST programs. There are five implementations of BLAST, three designed for nucleotide sequence queries (BLASTN, BLASTX, and TBLASTX) and two designed for protein sequence queries (BLASTP and TBLASTN) (Coulson, *Trends in Biotechnology*, 12: 76-80 (1994); Birren, et al., *Genome Analysis*, 1: 543-559 (1997)).

[0266] BLASTN takes a nucleotide sequence (the query sequence) and its reverse complement and searches them against a nucleotide sequence database. BLASTN was designed for speed, not maximum sensitivity, and may not find distantly related coding sequences. BLASTX takes a nucleotide sequence, translates it in three forward reading frames and three reverse complement reading frames, and then compares the six translations against a protein sequence database. BLASTX is useful for sensitive analysis of preliminary (single-pass) sequence data and is tolerant of sequencing errors (Gish and States, *Nature Genetics*, 3: 266-272 (1993), herein incorporated by reference). BLASTN and BLASTX may be used in concert for analyzing EST data (Coulson, *Trends in Biotechnology*, 12: 76-80 (1994); Birren et al., *Genuine Analysis*, 1: 543-559 (1997)).

[0267] Given a coding nucleotide sequence and the protein it encodes, it is often preferable to use the protein as the query sequence to search a database because of the greatly increased sensitivity to detect more subtle relationships. This is due to the larger alphabet of proteins (20 amino acids) compared with the alphabet of nucleotide sequences (4 bases), where it is far easier to obtain a match by chance. In addition, with nucleotide alignments, only a match (positive score) or a mismatch (negative score) is obtained, but with proteins, the presence of conservative amino acid substitutions can be taken into account. Here, a mismatch may yield a positive score if the non-identical residue has physical/chemical properties similar to the one it replaced. Various scoring matrices are used to supply the substitution scores of all possible amino