

as GCG program ProfileSearch) and Hidden Markov Models (HMMs)(Krough et al., *J. Mol. Biol.* 235:1501-1531 (1994); Eddy, *Current Opinion in Structural Biology* 6:361-365 (1996), both of which are herein incorporated by reference in their entirety). In both cases, a large number of common protein domains have been converted into profiles, as present in the PROSITE library, or HMM models, as in the Pfam protein domain library (Sonnhammer et al., *Proteins* 28:405-420 (1997), the entirety of which is herein incorporated by reference). Pfam contains more than 500 HMM models for enzymes, transcription factors, signal transduction molecules, and structural proteins. Protein databases can be queried with these profiles or HMM models, which will identify proteins containing the domain of interest. For example, HMMSW or HMMFS, two programs in a public domain package called HMMER (Sonnhammer et al., *Proteins* 28:405-420 (1997)) can be used.

[0271] PROSITE and BLOCKS represent collected families of protein motifs. Thus, searching these databases entails submitting a single sequence to determine whether or not that sequence is similar to the members of an established family. Programs working in the opposite direction compare a collection of sequences with individual entries in the protein databases. An example of such a program is the Motif Search Tool, or MoST (Tatusov et al. *Proc. Natl.*

Acad. Sci. 91: 12091-12095 (1994), the entirety of which is herein incorporated by reference.) On the basis of an aligned set of input sequences, a weight matrix is calculated by using one of four methods (selected by the user); a weight matrix is simply a representation, position by position in an alignment, of how likely a particular amino acid will appear. The calculated weight matrix is then used to search the databases. To increase sensitivity, newly found sequences are added to the original data set, the weight matrix is recalculated, and the search is performed again. This procedure continues until no new sequences are found.

[0272] Table 1 lists the nucleic acid molecules encoding homologs of known proteins.

Lengthy table referenced here

US20070050860A1-20070301-T00001

Please refer to the end of the specification for access instructions.

LENGTHY TABLE

The patent application contains a lengthy table section. A copy of the table is available in electronic form from the USPTO web site (<http://seqdata.uspto.gov/?pageRequest=docDetail&DocID=US20070050860A1>). An electronic copy of the table will also be available from the USPTO upon request and payment of the fee set forth in 37 CFR 1.19(b)(3).

SEQUENCE LISTING

The patent application contains a lengthy "Sequence Listing" section. A copy of the "Sequence Listing" is available in electronic form from the USPTO web site (<http://seqdata.uspto.gov/?pageRequest=docDetail&DocID=US20070050860A1>). An electronic copy of the "Sequence Listing" will also be available from the USPTO upon request and payment of the fee set forth in 37 CFR 1.19(b)(3).

1. A purified nucleic acid molecule which is or is complementary to a nucleotide sequence selected from the group consisting of SEQ ID NO:1684, SEQ ID NO:1685, SEQ ID NO:1686, and SEQ ID NO:1687.

2. The purified nucleic acid molecule according to claim 1, wherein said nucleic acid molecule encodes a fragment of a *D. v. virgifera* protein.

3. The purified nucleic acid molecule according to claim 2, wherein said *D. v. virgifera* protein is a fragment of a V-ATPASE.

4.-26. (canceled)

* * * * *