these regions, and including them can give rise to misleading
motif signals, particularly if common domains are found in
more than one protein in a set. Most importantly, this avoids
the detection of repetitive, purely structural patterns, such as
β-turns, coiled-coil heptads, or collagen repeats, because
these are unlikely to occur in the unstructured parts of
proteins that remain after this filtering. We also compare all
sequences in a set to each other and leave only one
representative of any homologous segments. We do this in
order to measure over-representation that is not the result of
homology; our assumption is that each of the remaining
instances of a particular motif has arisen convergently and is
thus an independent observation. We specifically avoid
removing regions of low complexity because linear motifs
frequently occur within them.

We then find all three to eight residue motifs in the
remaining sequence [9], and score their over-representation
as the binomial probability (P) of seeing them randomly in a
similar set of sequences (see Materials and Methods). This
allows multiple observations of an otherwise insignificant
motif to become statistically significant by over-representa-
tion, and readily accounts for sets of different sizes and
composition. For example, the SH3-binding pattern RxPxxP
readily occurs in about one out of 20 randomly selected
proteins, but its occurrence in seven sequences in a set of
nine becomes highly significant. We also compute P for all
closely related species based on whether or not the same
motifs are seen in the corresponding orthologues, and
multiply these to give a final score (Scons; see Materials and
Methods).

We applied our approach to interacting sets of proteins
from Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhab-
ditis elegans, and Homo sapiens [10–14]. For the first three
species, these datasets are from yeast two-hybrid screens;
human data comes from the human Proteome Resource
Database (HPRD) [14] and consists of hand-curated inter-
actions extracted from the literature (see Protocol S1). For
each dataset, we constructed a control by selecting random
sets of proteins of a similar length and number, and
performed the same calculations. We then defined a
benchmark

The Eukaryotic Linear Motif resource (ELM) [5] contains a
curated set of experimentally validated instances of binding
motifs (i.e., their location in a particular protein). This
provides several pertinent sets of proteins to test the
approach, namely each set of proteins containing a known
instance of a particular motif (e.g., all PxxP motif–containing
sequences known to interact with SH3 domains). Of 58
different sets, 22 contained at least four non-homologous
instances of the motif, and could be used to test our
approach. We ran the procedure on each set and monitored
where the known motif (or a variant) was found in the list of
all motifs ranked according to Scons. Despite many thousands
of possibilities, the approach detected the correct motif as
the very best ranked for 14 out of 22 and among the top ten
for an additional three (Table 1). Applying the confidence
threshold left eleven correct motifs at first rank, and no false
predictions (see legend to Table 1). Inspection showed that
those motifs that were either missed or scored poorly were