Self-interacting proteins, whose two or more copies can interact with each other, play important roles in cellular functions and the evolution of protein interaction networks (PINs). Knowing whether a protein can self-interact can contribute to and sometimes is crucial for the elucidation of its functions. Previous related research has mainly focused on the structures and functions of specific self-interacting proteins, whereas knowledge on their overall properties is limited. Meanwhile, the two current most common high throughput protein interaction assays have limited ability to detect self-interactions because of biological artifacts and design limitations, whereas the bioinformatic prediction method of self-interacting proteins is lacking. This study aims to systematically study and predict self-interacting proteins from an overall perspective. We find that compared with other proteins the self-interacting proteins in the structural aspect contain more domains; in the evolutionary aspect they tend to be conserved and ancient; in the functional aspect they are significantly enriched with enzyme genes, housekeeping genes, and drug targets, and in the topological aspect tend to occupy important positions in PINs. Furthermore, based on these features, after feature selection, we use logistic regression to integrate six representative features, including Gene Ontology term, domain, paralogous interactor, enzyme, model organism self-interacting protein, and betweenness centrality in the PIN, to develop a proteome-wide prediction model of self-interacting proteins. Using 5-fold cross-validation and an independent test, this model shows good performance. Finally, the prediction model is developed into a user-friendly web service SLIPPER (SeLf-Interacting Protein PrEdictor). Users may submit a list of proteins, and then SLIPPER will return the probability scores measuring their possibility to be self-interacting proteins and various related annotation information. This work helps us understand the role self-interacting proteins play in cellular functions from an overall perspective, and the constructed prediction model may contribute to the high throughput finding of self-interacting proteins and provide clues for elucidating their functions. Molecular & Cellular Proteomics 12: 10.1074/mcp.M112.021790, 1689–1700, 2013.

Self-interacting proteins are referred to as those whose two or more copies can interact with each other. Two copies of a protein interact with each other to form a homodimer, and three or more copies to form a homotrimer or higher order homo-oligomer.

Self-interacting proteins play important roles in cellular functions. The majority of the enzymes in the BRENDA database are self-interacting proteins (1), and the transition between different oligomeric states may regulate enzyme activity (2). Many multiprotein complexes such as proteasome, ribosome, and nucleosome contain homodimers (3). Moreover, the function of many channel proteins, which control the transport of small molecules and ions across cell membranes, relies on their homo-oligomers (2), and many transcriptional factors also function by binding DNA as a homodimer (2). In particular, protein self-interactions are crucial in cellular signal transduction. The oligomerization and activation of cell-surface receptors in response to an agonist usually are important steps during the transfer of a signal across the cell membrane. For example, growth hormone, tyrosine kinase, and G-protein-coupled receptors all use this approach (2).

The ability to self-interact can confer several different structural and functional advantages to proteins, including improved stability, allosteric regulation, control over the accessibility and specificity of active sites, as well as increased complexity (1–4). In addition, homo-oligomerization can also allow proteins to form large structures without increasing genome size and with increasing error control during synthesis (1, 4).

Besides focusing on the structures and functions of specific self-interacting proteins, recently, a few of their overall properties were elucidated. Ispolatov et al. (3) found that in the
protein interaction network (PIN)\(^1\) the average number of interaction partners of self-interacting proteins is two times higher than that of other proteins, and the likelihood of a protein to interact with itself is proportional to the number of its interaction partners. Self-interacting proteins were also found to have lower aggregation propensity, which is the self-interaction phenomenon of misfolded proteins potentially leading to detrimental effects on physiology (5). In 2010, Pérez-Bercoff et al. (6) found that genes encoding self-interacting proteins have higher duplication propensity.

Self-interacting proteins are also crucial for the evolution of PINs. Vázquez et al. (7) found that if self-interactions are considered the duplication-divergence model of network evolution will produce PINs with the clustering pattern similar to that observed in natural networks. This suggests the special role that self-interactions play in the emergence of network modularity. Recently, this viewpoint was further consolidated by Gibson et al. (8), Pereira-Leal et al. (9) also presented that the formation of many protein complexes composed of paralogous proteins results from the duplication of self-interacting proteins.

However, Gibson et al. (8) presented that currently the number of self-interacting proteins may be underestimated. Initially, they pointed that yeast two-hybrid and affinity purification with mass spectrometry, which are two most common high throughput assays used to detect protein interactions, have limited ability to discern protein self-interactions because of biological artifacts and design limitations. Furthermore, the examination of physical data (for example, proteins in PDB or enzymes in BRENDA) supports a higher proportion of self-interacting proteins than yeast two-hybrid and affinity purification with mass spectrometry studies indicate.

Most previous researches pay attention to the structures and functions of individual self-interacting proteins. Although there are a few related jobs as stated above, knowledge on their overall properties is still limited. Meanwhile, more importantly, under the condition of the limited self-interaction detection ability of two most common high throughput assays, to our current knowledge, there is no bioinformatic prediction model of self-interacting proteins.

Considering these issues, in this paper we systematically study and predict self-interacting proteins from an overall perspective. First, we systematically analyze the properties of self-interacting proteins from structural, functional, evolutionary, and network topological aspects. Second, we widely collect features that can efficiently discriminate self-interacting proteins from others. After feature selection, we employ a logistic regression framework to integrate multiple features to develop a self-interacting protein prediction model. This model can predict a probability score for each protein to measure its possibility to be a self-interacting protein. Finally, we develop this prediction model into a user-friendly web service SLIPPER (Self-Interacting Protein PrEdictor). For each submitted protein, SLIPPER can give the probability score and various related annotation information.

**MATERIALS AND METHODS**

**Integrated Human Protein Interaction Network and Self-interacting Protein Data**—To construct a human PIN as large as possible used to analyze the network topological property of self-interacting proteins and, at the same time, to obtain known human self-interacting proteins as comprehensive as possible, we integrated protein interaction data from multiple public databases. From the PDB (10) (provided by Gibson et al. (8)), DIP (version 20100614) (11), BioGRID (version 3.0.66) (12), IntAct (version 20100628) (13), and MINT (version 2010-05-05) (14) databases, we collected human protein interactions, including self-interactions supported by at least one piece of experimental evidence together with their standard Proteomics Standards Initiative-Molecular Interaction annotation information (15). The annotation information we collected included the interaction detection method, interaction type, and the publications where the interaction was shown. Here, we only focused on those interactions whose interaction types are annotated as “direct interaction” or its child types. Using SwissProt accession number (AC) (version Aug. 10, 2010) (16) to uniformly denote proteins, we finally obtained an integrated human PIN composed of 20,560 nonredundant interactions between 7153 proteins and then 1537 human self-interacting proteins (supplemental Table S1).

**Golden Standard Dataset**—To develop a self-interacting protein prediction model, first, the golden standard positive (GSP) and negative (GSN) datasets were determined. The GSP set was constructed based on the integrated human self-interacting protein datasets mentioned above. To ensure high quality data, the GSP set only contains self-interacting proteins as follows: (a) those reported by at least two publications or detected by at least two kinds of experimental technology; (b) those supported by crystallography structure data in PDB (10); or (c) those detected by at least one small scale experiment. Finally, 1198 high quality human self-interacting proteins constituted the GSP set. For the GSN set, considering the difficulty obtaining an experimentally negative dataset, we removed all known self-interacting proteins from the whole human proteome, and we then randomly chose 2000 proteins from the remaining human proteins to construct the GSN set. The GSN set was repeatedly constructed 100 times, and the means ± S.D. of the results are reported under “Results.” Here, to decrease the number of possible positives in the GSN set as much as possible, the removed known self-interacting proteins were not only 1537 ones of the “direct interaction” type mentioned above but also 2709 ones of more extensive “association” type that were collected from the same sources as described above (supplemental Table S1).

**Data Sources and Methods Related to Feature Collection**—Pfam domain (17) assignments of human proteins are from InterPro database (release 27.0) (18). Housekeeping genes are typically constitutive genes that are required for the maintenance of basic cellular functions and ubiquitously expressed in all tissues/cell types (19). The 6909 human housekeeping genes are provided by Zhu et al. (19), which are expressed in at least 16 of 18 tissues according to the expressed sequence tag (EST) data. Expression breadth is referred to as the number of unique tissues where a gene is expressed. The list of human

---

\(^1\) The abbreviations used are: PIN, protein interaction network; AC, accession number; AUC, area under the curve; DER, domain enrichment ratio; FP, false-positive; GER, Gene Ontology term enrichment ratio; GO, Gene Ontology; GSN, golden standard negative; GSP, golden standard positive; LR model, logistic regression model; LR, likelihood ratio; MI, mutual information; mRMR, minimum redundancy maximum relevance; ROC, receiver operating characteristic; TP, true positive; PDB, Protein Data Bank.
enzyme genes is from Ma et al. (20), and known human drug targets are
pased from the DrugBank database (version May, 2011) (21).

Self-interacting protein data of model organisms are collected from
PDB (10), DIP (version 20101010) (11), IntAct (version 20100628) (13),
and MINT (version 20100505) (14). Here, to improve the number of
proteins supported by model organism self-interacting proteins
we considered model organism self-interacting proteins not
only of “direct interaction” type but also of more extensive “associa-
tion” type. In total, we obtained 349 fly, 237 worm, and 1276 yeast
self-interacting proteins of association or its child interaction types
supported by at least one piece of experimental evidence. Pairwise
orthologs between human and model organisms were downloaded
from Inparanoid (release 7) (22). Gene Ontology (GO) annotations of
human proteins were obtained from GO project website (downloaded
on Sept. 9, 2011) (23).

The evolutionary rate of a protein is defined as the ratio of
the number of nonsynonymous substitutions per nonsynonymous site
(dN) to the number of synonymous substitutions per synonymous site
to the number of synonymous substitutions per synonymous site
(S) (21). We followed the procedure described previously (24) to
calculate an evolutionary rate of human proteins, adopting Pan trogl-
dytes as reference species.

The data on age assignments of human proteins are also from
our previous work (24). Human proteins were assigned to 16 age classes,
from the youngest “Homo sapiens” class containing proteins only
found in H. sapiens to the most ancient “Cellular organisms” class
containing those that originated in the common ancestor of three
domains of tree of life (Eukaryota, Bacteria, and Archaea) (Fig. 1C)
(see more detail in Ref. 24).

The degree of a given node in a network is defined as the number
of interaction partners of this node in the network. Betweenness
centrality of node n in a network is computed as shown in Equation 1,

\[
B(n) = \frac{2 \times \sum_{s \neq t \neq n} \sigma_{st}(n)}{(N - 1)(N - 2)}
\]

where \(\sigma_{st}\) denotes the number of shortest paths between node s and
node t in the network; \(\sigma_{st}(n)\) denotes the number of shortest paths
between node s and t which go through node n, and N is the total
number of nodes in the network. The numerator in Equation 2

\[
\sum_{s \neq t \neq n} \sigma_{st}(n)
\]

characterizes the total influence of node n on the “information trans-
fer” between all other node pairs in the network, which is normalized
by its maximum possible value \((N - 1)(N - 2)/2\), i.e. the total number
of node pairs (excluding node n) in the network (25).

Here, pairwise paralogous proteins in a species were identified as
those of which the BLAST E values of the amino acid sequences are
not larger than 10\(^{-2}\). We set a relatively loose E-value cutoff

to paralog identification, mainly to obtain a relatively large coverage
of proteins having paralogous interactors. The amino acid sequences
of human proteins were downloaded from SwissProt on Aug 10, 2010
(16). We calculated the sequence similarity of a pair of paralogous
proteins as the product of identity, query sequence coverage, and
subject sequence coverage in the BLAST sequence alignment of this
pair of proteins, where identity is the percentage of identical amino
acids within the BLAST alignment interval, and query/subject se-
quence coverage is the number of identical amino acids in the align-
ment interval divided by the total number of amino acids of query/subject
sequence.

Domain enrichment ratio (DER) is used to measure the enrich-
ment of a domain in the GSP dataset, which is computed as the probability
of observing this domain in the GSP dataset divided by that in the
whole human proteome. We computed the GO term enrichment ratio
(GER) in a similar way.

Minimum Redundancy Maximum Relevance (mRMR) Feature Se-
lection and Mutual Information (MI) — Here, mRMR method is used for
feature selection (26, 27). mRMR ranks each feature according to its
relevance to the classification variable and redundancy to other fea-
tures. Both relevance and redundancy are measured based on MI. MI
denoted by \(I\) between two discrete random variables \(X\) and \(Y\) is
defined as shown in Equation 3,

\[
I(X,Y) = \sum_{x,y} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) \tag{Eq. 3}
\]

where \(p(x,y)\) is the joint probabilistic distribution function, and \(p(x)\) and
\(p(y)\) are the respective marginal probability distribution functions (26).
MI quantifies the mutual dependence between two random variables,
i.e. it measures how much knowing one variable reduces uncertainty
about the other. It is zero if and only if \(X\) and \(Y\) are independent; it
increases as their mutual independence decreases, and it reaches the
other extreme when \(X\) and \(Y\) totally share the information (for example
\(X\) and \(Y\) are identical) (28). Using MI, the relevance \(D\) of a candidate
feature \(f_{\text{candidate}}\) to classification variable \(c\) and its redundancy \(R\) with
other features are defined as shown in Equation 4,

\[
D = l(f_{\text{candidate}},c) \quad \text{and} \quad R = \frac{1}{|S|} \sum_{f \in S \setminus \{f_{\text{candidate}}\}} l(f,c) \tag{Eq. 4}
\]

Suppose \(S\) and \(T\) are, respectively, the already-selected and to-
be-selected feature sets, and \(|S|\) and \(|T|\) are the number of features in \(S\)
and \(T\). In the mRMR method, the most relevant feature to classification
variable is firstly selected. The remaining features in \(T\) are moved
into set \(S\) one by one, requiring each time the selected feature
\(f_j\) to optimize Equation 5 (27)

\[
\max_{f \in T} \left( l(f,c) - \frac{1}{|S|} \sum_{f \in S} l(f,c) \right) \tag{Eq. 5}
\]

The mRMR program was downloaded from its website.

Logistic Regression Model (LR Model)— Denote the classification
variable by \(c\) and the features by \(f_1, f_2, …, f_n\), and let \(c = 1\) when
a protein can interact with itself and \(c = 0\) when it cannot. The
logistic regression model (LR model) used to predict the protein
classification based on features \(f_1, f_2, …, f_n\) is expressed as shown in Equation 6,

\[
P(c = 1) = \frac{1}{1 + e^{-\beta_0 + \beta_1 f_1 + \beta_2 f_2 + \cdots + \beta_n f_n}} \tag{Eq. 6}
\]

where \(\beta_0\) and \(\beta_1, \beta_2, …, \beta_n\) are, respectively, constant and regression
coefficients of features \(f_1, f_2, …, f_n\) and the \(P(c = 1)\) denoted by the
probability score in this work represents the possibility of a protein
to be a self-interacting protein (29). Here we use R software to
train the LR model based on the gold standard dataset (30).

Receiver Operating Characteristic (ROC) Curve and 5-Fold Cross-
validation — To evaluate the performance of the prediction model, the
ROC curve together with 5-fold cross-validation protocol is used (31).
The ROC curve can show the efficacy of a dichotomous classifier by
mapping sensitivity to specificity at varying cutoffs. Sensitivity (TP/TPR
and specificity (1- FP/TP) are, respectively, referred to as a correctly
classified fraction in the positive test set (T) and in the negative test
set (F), where TP is the predicted positives in T (that is true positives),
and FP is the predicted positives in F (that is false positives). Here,
Study and Prediction of Self-interacting Proteins

TABLE I

| Property                     | The average value (± S.D.) | p value (rank sum test) |
|------------------------------|---------------------------|-------------------------|
|                              | Self-interacting proteins | Other proteins          |                        |
| Domain number                | 1.3643 (± 1.1924)         | 0.8240 (± 1.0446)       | <10⁻⁴                  |
| Evolutionary rate            | 0.2472 (± 0.3382)         | 0.3665 (± 0.7182)       | <10⁻⁴                  |
| Age (mean rank)              | 9624.96                   | 8091.38                 | <10⁻⁴                  |
| Degree                       | 10.9165 (± 18.3772)       | 4.4878 (± 7.6432)       | <10⁻⁴                  |
| Betweenness centrality       | 0.0013 (± 0.0038)         | 0.0003 (± 0.0011)       | <10⁻⁴                  |

*Here we count the domain number of each protein excluding domain repeats. Actually, we obtain the same conclusion when considering domain repeats (average domain number for self-interacting proteins and other proteins: 2.0403 > 1.1987, rank sum test, p < 10⁻⁴).

*Age is an ordered categorical variable, and thus here we show mean rank instead of mean value.

RESULTS

Properties of Self-interacting Proteins—To understand the role that self-interacting proteins play in cellular functions from an overall perspective, we systematically analyzed their properties from multiple aspects.

First, to obtain a list of known human self-interacting proteins as comprehensively as possible, we collected human self-interaction proteins supported by at least one piece of experimental evidence from multiple publicly available resources (see "Materials and Methods"). In total, 1537 non-redundant known human self-interacting proteins were obtained (supplemental Table S1).

Based on this integrated dataset, we studied the properties of self-interacting proteins from the following aspects. First, we compared the average domain number of self-interacting proteins with that of other proteins. Protein domains are the basic structural and functional units of a protein, and the domain number generally can reflect the complexity of a protein's structure and function (33). We find that, compared with other proteins, self-interacting proteins contain significantly more domains, indicating they tend to have a more complex structure and function (Table I). Then, by computing evolutionary rates and assessing original ages of proteins, we observe that self-interacting proteins tend to be evolutionarily conserved and ancient (Table I). Evolutionary conservation usually implies functional importance. The result on original age is also consistent with two previous studies based on the three-dimensional structural data of yeast proteins (34) and a small scale yeast homodimeric protein dataset (35), which both showed that ancient proteins are highly enriched with homodimeric proteins. Furthermore, self-interacting proteins are shown to be significantly enriched with enzyme genes and housekeeping genes, which are just consistent with our previous knowledge on the importance of homo-oligomerization in functional implementation of enzymes and more extensive cellular functions (Table II) (2). Surprisingly, about 40% of known human self-interacting proteins we collect are known drug targets, which is several times higher than that of other proteins (Table II). Finally, in the topological aspect of the network, we analyzed the degree and betweenness centrality of self-interacting proteins in PINs. These two topological indices with different emphases can both be used to quantify the topological importance of individual nodes in the network.

Degree is the number of direct neighbors of a given node, capturing the information of a node’s local importance. In contrast, betweenness centrality is related to the frequency with which a node appears on all the shortest paths between all pairs of nodes in the network (see "Materials and Methods"), thus not only considering local information but also reflecting a semi-global manner of how a node influences the "communication" between other nodes in the network (25, 36). Self-interacting proteins are found to have a significantly higher degree and betweenness centrality than other proteins in the PIN (Table I), suggesting the crucial role self-interacting proteins play in PINs. The result on a higher degree of self-interacting proteins is also consistent with a previous study (3). The higher degree of self-interacting proteins in PINs might be related to gene duplication. Different from a non-self-interacting protein, the duplication of a self-interacting protein will lead to an additive interaction partner (paralogous interactor) because of the conservation of self-interaction (8). Therefore, as proteins continuously duplicate, self-interacting proteins will accumulate more interaction partners than other proteins.

These results in structural, evolutionary, functional, and network topological aspects all indicate the importance of self-interacting proteins in cellular functions, and they also suggest the potential special role they play in drug research. In addition, we also obtain similar results in yeast, indicating these conclusions are universal across different organisms (supplemental Tables S2 and S3).

Features Used to Predict Self-interacting Proteins—To construct a self-interacting protein prediction system, features...
were first collected, which can efficiently discriminate between self-interacting proteins and other proteins.

From the analyses above, we found that on multiple properties the self-interacting proteins are significantly different from other proteins. Theoretically, all these features can be used to predict self-interacting proteins. We use likelihood ratio (LR) to measure the discrimination power or confidence level of feature f, which is defined as the ratio of the probability of feature f observed in GSP set to that in GSN set (31). In theory, LR(f) > 1 means that feature f has the ability to distinguish self-interacting proteins from other proteins. As shown in Fig. 1, A–G, as we expected all these features have the ability to predict self-interacting proteins. Furthermore, different bins of the same feature have different prediction abilities, and the tendency is in agreement with the results in Tables I and II for all these features. For example, for feature “degree in the PIN,” proteins with different degrees have different confidence levels to be a self-interacting protein. The higher the degree of a protein is, the more possibly this protein can interact with itself. This tendency is consistent with the previous result that self-interacting proteins tend to have higher degree in the PIN (Table I). Besides the features above, we also consider another four features as shown below.

Model Organism Self-interacting Protein—This feature was successfully used previously to predict protein interactions (not including protein self-interactions) (31, 37). Here, we introduce it to predict protein self-interactions. Based on interaction conservation across different organisms (38, 39), if a human protein’s ortholog in another model organism can interact with itself, this human protein could be a self-interacting protein. To predict human self-interacting proteins, we collected self-interacting proteins of three model organisms, including fly, worm, and yeast. We see that the LRs of bins supported by model organism self-interacting proteins are larger than 1, and a protein supported by self-interacting proteins in more model organisms has a higher confidence level to self-interact, indicating the prediction efficacy of this feature (Fig. 1H).

DER and GER—It is thought that self-interacting proteins tend to contain special domains or have special biological functions, and hence novel self-interacting proteins could be predicted by identifying enriched domains or GO functional terms among known self-interacting proteins. To test this logic, the GSP set is randomly divided into three equal subsets, two are used to define enriched domains/GO terms and compute corresponding DERs/GERs (see “Materials and Methods”), and the remaining one is used to test the prediction ability of the enriched domains/GO terms. This process was repeated three times, and the LRs were averaged. We observed a clear correlation between LR and DER/GER, suggesting that these two features are suitable for self-interacting protein prediction (Fig. 1, I and J). A complete list of enriched domains/GO terms in the GSP set is available in supplemental Table S4. We speculate that many of these enriched domains can self-interact, and thus proteins that contain such domains can self-interact by means of domain-domain self-interactions. An example is the “S-100/ICaBP-type calcium-binding domain” (Pfam AC: PF01023, DER = 5.31). As shown by PDB structure 1BT6, the self-interaction of protein “S100-A10” (SwissProt AC: P60903) in the GSP set is mediated by this domain’s self-interaction. As for the reason of GER’s effectiveness for self-interacting protein prediction, we think that implementation of the many functions relies on the homologization ability of proteins, and therefore a protein with such function could be a self-interacting protein. For example, for the GO term “glutathione transferase activity” (GO: 0004364), there are 13 proteins with this molecular function in the GSP set, which is significantly larger than random expectation (GER = 7.46), suggesting the importance of homodimerization for glutathione transferase activity. Therefore, a protein with glutathione transferase activity has a high probability to be a homodimer.

Paralogous Interactor—When self-interacting proteins duplicate, inheritance of self-interactions will result in the interactions between paralogous proteins. It has been proposed that most current interactions between paralogs are derived from the duplication of ancestral self-interacting proteins (3). Therefore, we speculate that a protein having paralogous interactors (i.e., interaction partners) may still maintain its ancestor’s self-interaction ability. The probability of this protein to be a self-interacting protein could be correlated to its sequence similarity with its paralogous interactors. The higher the protein’s sequence similarity is with its paralogous protein, then the higher its sequence similarity is with its common ancestral protein, and therefore, the more the protein possibly maintains the ancestor’s self-interaction ability. The LR result validates these speculations (Fig. 1K).

TABLE II
The fraction of proteins belonging to a certain protein class among self-interacting proteins and that among other proteins

| Property                | The fraction of proteins belonging to a certain protein class (%) | p value ($\chi^2$ test) |
|-------------------------|---------------------------------------------------------------|-------------------------|
|                         | Self-interacting proteins | Other proteins          |
| Enzyme                  | 26.9                        | 9.8               | $<10^{-4}$                |
| Housekeeping gene       | 56.0                        | 31.6              | $<10^{-4}$                |
| Drug target             | 39.6                        | 8.2               | $<10^{-4}$                |

Ranking Features and Feature Selection—By computing LRs, we confirmed that all the 11 features can be used to predict self-interacting proteins. However, apparently their prediction abilities were different. Meanwhile, they are not independent with each other (supplemental Table S5). It has been recognized that the combination of all individually efficient features does not necessarily produce the best prediction performance (40). Often low relevant or high redundant features have no positive impact and even sometimes have
FIG. 1. LRs of various features used to predict self-interacting proteins. LR of feature $f$ is calculated as the ratio between the fraction of proteins with feature $f$ in the GSP set and that in the GSN set. A, domain number. B and C, evolutionary rate and age (see “Materials and Methods”). The evolutionary rate/age of proteins in the first bin cannot be estimated. D, housekeeping gene. Here, housekeeping genes are referred to as those expressed in at least 16 of 18 tissues in human (see “Materials and Methods”). Expression breadth is the number of tissues where a gene is expressed. The housekeeping genes are further grouped into three bins based on the expression breadth (H2 to H4). E, enzyme. F and G, degree and betweenness centrality in the PIN. The first bin of each of these two features is composed of those proteins that are not in the PIN and hence have no degree/betweenness centrality. H, model organism self-interacting protein. Not supported by model organism means that a protein’s orthologs in any model organism (fly, worm, and yeast) are not known self-interacting proteins. I and J, GER and DER. If a protein contains multiple enriched GO terms/domains, the largest one of the corresponding GERs/DERs is used. K, paralogous interactor. Proteins having paralogous interaction partners are further binned based on their sequence similarity and BLAST E-value with their paralogous interactors (see “Materials and Methods”). When a protein has multiple paralogous interactors, the largest sequence similarity and the smallest E-value are used. GSN is repeatedly constructed 100 times (see “Materials and Methods”), and thus error bars in these figures represent SDs of LRs of 100 times.
negative impact on the classification performance, simultaneously leading to the decrease of generalization capability of prediction model and resource waste. Therefore, before integrating multiple features to construct the prediction model, feature selection is necessary.

First, based on MI, we compared the discriminative power of these features, i.e. their relevance to the classification variable. MI quantifies the mutual dependence between two random variables, i.e. it measures how much knowing one variable reduces uncertainty about the other (see “Materials and Methods”). Therefore, MI between feature \( f \) and protein classification (self-interacting protein class and nonself-interacting protein class) by and large could reflect the ability of feature \( f \) to predict self-interacting proteins. Because multiple features as well as protein classification are categorical variables, when computing MI we convert continuous variables into discrete variables by binning shown in Fig. 1. As a result, we find that the features of GER, degree, and betweenness centrality in the PIN have relatively high self-interacting protein prediction power, followed by paralogous_interactor and DER, although two evolutionary features and “model organism self-interacting protein” have low discrimination ability (Table III).

Besides the relevance of features to the target classification, their mutual redundancy is also an important factor that should be considered in feature selection. Here, we used mRMR feature selection method that ranks features based on both relevance and redundancy. Starting from the most relevant feature, mRMR sequentially selects a feature from the to-be-selected feature set, which maximizes its relevance to the classification variable and simultaneously minimizes its redundancy with features in the already-selected feature set (see “Materials and Methods”). We apply mRMR to these 11 features. We see that the features ranked at the bottom are either those with low relevance to the protein classification (such as evolutionary_rate and age) or those relatively high redundant with the features ranked in front (for example, as shown by supplemental Table S5, degree has high correlation with betweenness_centrality, and domain_number has high correlation with DER) (Table IV).

**Self-interacting Protein Prediction Model Based on Logistic Regression**—We use logistic regression model (LR model) to integrate multiple features to construct the self-interacting protein prediction model. LR model previously has been successfully used to predict protein (nonself) interactions (29, 41). It can integrate various heterogeneous data of discrete or continuous type and is robust against redundancy between features to some extent (42). Meanwhile, it provides a comparable, normalized probability_score for each interested protein, which measures its possibility to be a self-interacting protein (see “Materials and Methods”). The ROC curve is used to assess the performance of the prediction model. Each point on the ROC curve denotes the model’s sensitivity and specificity against a particular probability_score cutoff, and thus the ROC curve shows the tradeoff between sensitivity and specificity (see under “Materials and Methods”). A good prediction model has a ROC curve climbing rapidly toward the top left corner of the graph, which can be quantified by the area under the curve (AUC). AUC = 1.0 indicates a perfect classifier, whereas AUC = 0.5 means a noninformative prediction. The closer the AUC is to 1.0, the more efficient the classifier is (29).

Using the LR model, we first constructed a classifier for each feature. Assessment based on a 5-fold cross-validation protocol.
shows that these single-feature prediction models have different prediction abilities (Fig. 2). The order of these features according to AUCs is by and large consistent with that according to their MI with protein classification shown in Table III.

Furthermore, we constructed prediction models integrating multiple features. Starting from the most efficient single-feature model of GER, according to the order of features given by mRMR (Table IV), we added the features one by one into the prediction model. Ultimately 11 models were constructed, from the one only composed of feature GER (Model_1) to the one integrating all the 11 features (Model_11). As shown in Fig. 3, we first observe that the performance of all the prediction models integrating multiple features is better than that of any single-feature model. Furthermore, as the features are added into the model one by one, in the beginning stage the efficacy of the model gradually increases; however, after the number of features exceeds 6, the AUC does not rise any more but slightly declines. This indicates that the more features integrated into the classifier, the better the performance of the classifier is not necessary to be. We can obtain good classification efficacy only by integrating “representative” features with relatively high relevance and low mutual redundancy. The self-interacting protein prediction model integrating six features, including GER, paralogous interactor, betweenness centrality in the PIN, enzyme, DER, and model organism self-interacting protein (Model_6), has the highest AUC of 0.884, suggesting that it has strong ability to predict self-interacting proteins (Fig. 3).

Besides 5-fold cross-validation, we also use an independent test set to assess the performance of Model_6. Previously, in total we collect 1537 known human self-interaction proteins, from which we choose 1198 high quality ones to

![Fig. 2. ROC curves of single-feature prediction models of self-interacting proteins based on the 5-fold cross-validation.](image)

**Fig. 2.** ROC curves of single-feature prediction models of self-interacting proteins based on the 5-fold cross-validation. Different models are distinguished by different colors. GSN is repeatedly constructed 100 times (see “Materials and Methods”), and thus the presented ROC AUC in the figure is mean ± S.D. of the results of 100 times. The features are ordered according to the decreasing AUC’s average values of 100 times. The ROC curves in the figure are drawn based on GSP and a random GSN.

![Fig. 3. ROC curves of self-interacting protein prediction models integrating n features based on 5-fold cross-validation.](image)

**Fig. 3.** ROC curves of self-interacting protein prediction models integrating n features based on 5-fold cross-validation. Starting from the single-feature model of GER (Model_1), the features are integrated into the model one by one according to the order given by mRMR shown in Table IV. In total, 11 models are constructed, from Model_1 to Model_11 integrating all the 11 features. Their ROC curves are plotted by different colors. GSN is repeatedly constructed 100 times (see “Materials and Methods”), and thus the presented AUCs of the 11 models in the figure are the mean ± S.D. of the results of 100 times. The ROC curves in the figure are drawn based on GSP and a random GSN.

![Fig. 4. Average probability score of known human self-interacting proteins removing the GSP set (D1) and other proteins (that is the whole human proteome removing the GSP set and D1) (D2).](image)

**Fig. 4.** Average probability score of known human self-interacting proteins removing the GSP set (D1) and other proteins (that is the whole human proteome removing the GSP set and D1) (D2). The result is based on Model_6, which is constructed based on GSP and a random GSN from 100 repeatedly constructed ones (see “Materials and Methods”).
construct the GSP set (see “Materials and Methods”). The remaining 339 proteins can be viewed as an independent test set. We find that the average probability_score of the proteins in this test set given by Model_6 is 0.60, which is significantly higher than that of other proteins (rank sum test, p < 10^{-4}) (Fig. 4). This result further indicates the power of Model_6 to discriminate self-interacting proteins from others.

In addition, besides human, we also applied this prediction model integrating six features to yeast, and we obtained similarly good prediction performance (supplemental Figs. S1 and S2).

Web Service for the Self-interacting Protein Prediction—We developed the self-interacting protein prediction model integrating six features (Model_6) into a user-friendly web service SLIPPER, which can be assessed on line. Users may submit a group of proteins (Fig. 5A), and then SLIPPER will return the probability_scores measuring their possibility to be self-interacting proteins (Fig. 5B) and various related annotation information such as interaction partners in the PIN, domains, evolutionary rate, and so on, facilitating their further functional researches (Fig. 5C).

By SLIPPER, we find that some proteins with high probability_scores that are not in the known human self-interacting protein dataset integrated by us have already been experimentally validated to be a self-interacting protein. For example, transcriptional factor HNF4A (SwissProt AC: P41235, probability_score = 0.8826) has been found early to bind DNA as a homodimer. Type 1 maturity-onset diabetes of the young (MODY1) is just caused by HNF4A homodimerization failure (43). Stoffel et al. (43) found that because of a nonsense mutation in exon 7, HNF4A fails to homodimerize and bind DNA, and it loses the transcriptional transactivation activity. This leads to abnormal expression of downstream genes involved in glucose transport and glycolysis, further causes the defect of glucose-stimulated insulin secretion, and ultimately results in diabetes. As another example, DCC (SwissProt AC: P43146, probability_score = 0.7936) homodimerization induced by netrin has been found to play a key role in the axon guidance (44). These examples further indicate that our prediction system works well and can provide clues for elucidating protein functions.

DISCUSSION

Proteins that can interact with their copies, play important roles in cellular functions and the evolution of PINs. Here, we systematically elucidate the overall properties of this class of proteins from multiple aspects. Meanwhile, based on these properties, we successfully develop the first bioinformatic prediction model of self-interacting proteins, which achieves good performance by cross-validation and an independent test. Finally, the constructed prediction model is developed into a user-friendly web service, which can give probability_scores and abundant annotations for proteins submitted by users.

All features that self-interacting proteins are found to have (Tables I and II and Fig. 1) are mutually consistent, i.e. two features that are both positively correlated with protein classification (Fig. 1) also have large or small positive correlation coefficients between each other (supplemental Table S5), and two that are respectively positively and negatively correlated with protein classification (Fig. 1) have strong or weak negative correlation with each other (supplemental Table S5). Many correlation relationships between pairwise features have been reported previously. For example, Yang et al. (33) showed that proteins without known domains evolve faster than other proteins, and single-domain proteins evolve faster than multidomain proteins. This suggests negative correlation between domain number and evolutionary rate. In addition to the negative correlation between evolutionary rate and degree or betweenness centrality in the PIN (45, 46), positive correlation between age and degree in the PIN (25, 47), and the evolutionary conservation of housekeeping genes (48) have also been reported. Especially Hase et al. (49) found that drug target molecules have a higher average degree than others in the PIN, and thus the feature that self-interacting proteins are significantly enriched with drug targets is also consistent with the feature that they have higher degree in the PIN.

The constructed self-interacting protein prediction model integrating six features has several advantages. First, unlike some other machine learning methods that can only give the judgment of yes or no (for example SVM), the prediction model based on logistic regression can predict a mutually comparable probability_score for each protein, which quantifies its possibility to be a self-interacting protein. Second, the prediction coverage of our model equals 1. Profiting from the strategy of feature binning, in our model we can also make use of the information of “the zeroth bin” (for example, for feature betweenness centrality in the PIN, “the zeroth bin” is composed of proteins not in the PIN and thus without betweenness centrality (Fig. 1G)), which we think is also valuable for the judgment of a protein’s self-interaction possibility. Therefore, our model is able to assign a probability_score for...
each protein in the proteome, even though the protein belongs to the zeroth bin. In addition, our model is economical and efficient. Attributed to feature selection, the prediction model achieves good performance only by integrating several representative features with relatively high relevance and low mutual redundancy.

In our work, the GSN set is repeatedly constructed 100 times (see “Materials and Methods”), and the mean ± S.D. of the results of 100 times is reported. From “Results,” we see that actually the effect of this randomness in the construction of GSN on our results is small. Meanwhile, we also find that the ratio between the size of GSP and that of GSN does not have significant effect on our results (supplemental Table S6).

In addition, we also observe that GSP tends to have a higher “well studied” degree than GSN; however, we think it is more probable that such a higher well studied degree of self-interacting proteins is the result of their biological importance. Even though we balance the “well studied” degree of GSP and GSN, the special features (suggesting the biological importance) of self-interacting proteins still hold, and the prediction model integrating these features still have the ability to discriminate self-interacting proteins from other proteins (supplemental Tables S7–S12). What’s more, results from high throughput interaction datasets also suggest that real self-interacting proteins just tend to have a higher well studied degree than other proteins (supplemental Table S13).

Of course, we admit that our prediction model, as a bioinformatics method, heavily depends on the currently available related biological data. With the related biological data continuously updated, will the main results in this paper still hold? Taking protein interaction data as an example, during the revision of the manuscript, we updated the human self-interacting protein prediction model based on the latest protein interaction data together with known self-interacting protein data, and we found that all main results and conclusions in our work stay unchanged (see supplemental Tables S14 and S15, and supplemental Figs. S3 and S4 for the results). By SLIPPER, we will periodically update and perfect our prediction model based on continuously updated and perfected known self-interacting protein data and various other related data. Now the updated prediction model of human has been available in SLIPPER, by selecting “human v20121201” in the “Select Organism and Prediction_model_version” menu on the homepage of SLIPPER (Fig. 5A).

In summary, the elucidation of the properties of self-interacting proteins in structural, functional, evolutionary, and network topological aspects helps us understand their roles in cellular functions from an overall perspective. Meanwhile, under the condition of the limited ability to detect self-interactions of the two current most common high throughput protein interaction assays, our self-interacting protein prediction model provides an efficient complementary approach that contributes to the high throughput finding of self-interacting proteins. The developed web service SLIPPER for self-interacting protein prediction also provides a useful tool. Researchers working on a particular protein may obtain its probability_score to be a self-interacting protein by SLIPPER, and the experimental validation of a protein with a high score should be preferentially considered. Knowing whether a protein can self-interact, can contribute to and sometimes is crucial for the elucidation of protein functions. Meanwhile, various annotations of proteins presented by SLIPPER will also provide more clues for deep functional research.

Acknowledgments—We thank Haiyuan Yu, Xiaohui Wang, Lijun Tang, Dong Yang, Ning Li, and HaoGuo for fruitful discussions. Author Contributions: F.H. and D.L. provided guidance and revised the manuscript. Z.L. designed the study, carried out the study, and wrote the manuscript. J.Z., J.W., and L.L. participated in the analyses. D.L. and F.G. developed the web service SLIPPER.

* This work was supported by National Natural Science Foundation of China Grant 31271407, Chinese National Key Program of Basic Research Grants 2012CB910300 and 2011CB910202, Chinese High Technology Research and Development Grant 2012AA020201, National Key Technology R&D Program Grant 2012BAI29B07, and State Key Laboratory of Proteomics Grant SKLP-Q201106.

** This article contains supplemental material.

** To whom correspondence may be addressed: Dong Li, Tel.: 86-010-6817777-1155; Fax: 86-010-80705225; E-mail: lidong.bprc@foxmail.com, or Fuchu He, Tel.: 86-010-68171208; Fax: 86-010-68177417; E-mail: hefnc@nic.bmi.ac.cn.

REFERENCES
1. Hashimoto, K., Niahi, H., Bryant, S., and Panchenko, A. R. (2011) Caught in self-interaction: evolutionary and functional mechanisms of protein homooligomerization. Phys. Biol. 8, 035007
2. Marianayagam, N. J., Sunde, M., and Matthews, J. M. (2004) The power of two: protein dimerization in biology. Trends Biochem. Sci. 29, 618–625
3. Isolatov, I., Yuryev, A., Mazo, I., and Maslov, S. (2005) Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res. 33, 3629–3635
4. Goodsell, D. S., and Olson, A. J. (2000) Structural symmetry and protein function. Nucleic Acids Res. 28, 105–153
5. Chen, Y., and Dokholyan, N. V. (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. Mol. Biol. Evol. 25, 1530–1533
6. Pérez-Bercoff, A., Makino, T., and McLyasgath, A. (2010) Duplicability of self-interacting human genes. BMC Evol. Biol. 10, 160
7. Vázquezza, A., Flaminia, A., Maritana, A., and Vespignani, A. (2003) Modeling of protein interaction networks. ComPlexUs 1, 38–44
8. Gibson, T. A., and Goldberg, D. S. (2009) Questioning the ubiquity of neofunctionalization. PLoS Comput. Biol. 5, e1000252
9. Pereira-Leal, J. B., Levy, E. D., Kamp, C., and Teichmann, S. A. (2007) Evolution of protein complexes by duplication of homeric interactions. Genome Biol. 8, R51
10. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235–242
11. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 32, D449–D451
12. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz A., and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34, D535–D539
13. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssens, M., Khens, L., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, J., and Hermjakob, H. (2010) The IntAct molecular interaction database in 2010.
29. Lin, N., Wu, B., Jansen, R., Gerstein, M., and Zhao, H. (2004) Information assessment on predicting protein-protein interactions. BMC Bioinformatics 5, 154

30. R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

31. Li, D., Liu, W., Liu, Z., Wang, J., Liu, Q., Zhu, Y., and He, F. (2008) PRINCESS, a protein interaction confidence evaluation system with multiple data sources. Mol. Cell. Proteomics 7, 1043–1052

32. SPSS, Inc. (1999) SPSS Base 10.0, User’s Guide, pp. 431–434, SPSS, Inc., Chicago

33. Yang, D., Zhong, F., Li, D., Liu, Z., Wei, H., Jiang, Y., and He, F. (2012) General trends in the utilization of structural factors contributing to biological complexity. Mol. Biol. Evol. 29, 1957–1968

34. Kim, W. K., Henschel, A., Winter, C., and Schroeder, M. (2006) The many faces of protein-protein interactions: A compendium of interface geometry. PLoS Comput. Biol. 2, e124

35. Kim, W. K., and Marcotte, E. M. (2008) Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. PLoS Comput. Biol. 4, e1000232

36. Estrada, E. (2006) Protein bipartivity and essentiality in the yeast protein-protein interaction network. J. Proteome Res. 5, 2177–2184

37. Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavison, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005) Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol. 23, 951–959

38. Pagel, P., Mewes, H. W., and Frishman, D. (2004) Conservation of protein-protein interactions–lessons from ascomycota. Trends Genet. 20, 72–76

39. Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R., and Ideker, T. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. U.S.A. 100, 11394–11399

40. Poch, I., Long, F., and Ding, C. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238

41. Xia, Y., Lu, L. J., and Gerstein, M. (2006) Integrated prediction of the helical membrane protein interactome in yeast. J. Mol. Biol. 357, 339–349

42. Xia, Y., Franzosa, E. A., and Gerstein, M. B. (2009) Integrated assessment of genomic correlates of protein evolutionary rate. PLoS Comput. Biol. 5, e1000393

43. Stoffel, M., and Duncan, S. A. (1997) The maturity-onset diabetes of the young (MODY1) transcription factor HNF4α regulates expression of genes required for glucose transport and metabolism. Proc. Natl. Acad. Sci. U.S.A. 94, 13209–13214

44. Arakawa, H. (2004) Ntrin-1 and its receptors in tumorigenesis. Nat. Rev. Cancer 4, 978–987

45. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002) Evolutionary rate in the protein interaction network. Science 296, 750–752

46. Hahn, M. W., and Kern, A. D. (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. 22, 803–806

47. Eisenberg, E., and Levanon, E. Y. (2003) Preferential attachment in the evolution of protein interaction networks and their implications on drug design. PLoS Comput. Biol. 5, e1000550