convex (SA_6 = very-low) or strongly hydrophob (HY_PHOB = very-high) segments of the surface are likely to be docking sites. On the other hand, if a surface segment is very concave (SA_6 = very-high) or plane and not strongly hydrophob (SA_6 = medium $\wedge$ HY_PHOB = low) or convex and not strongly hydrophob (SA_6 = low $\wedge$ HY_PHOB = low) then it usually is not a docking site. Thus, different types of docking and non-docking sites were discovererd.

The classification algorithm always chose SA_6 as the first attribute and typically chose HY_PHOB as the second one. We conclude that SA_6 and HY_PHOB include the most information (w.r.t. the features investigated) for the classification of docking sites.

Conclusions

We have developed an efficient classification algorithm for large databases. The algorithm was used to classify protein docking sites based on a set of geometric and physico-chemical surface features. Although only a small set of features was investigated, our experiments yielded some interesting and useful classification rules. More features should be considered to increase the confidence of the rules. The following features seem to be potentially useful: carbon, oxygen, and sulphur content, hydrogen-ion-donor or acceptor property and aminoacid content in a defined neighborhood of a surface point.

References

[AHPS 95] Ackermann F., Herrmann G., Posch S., Sagerer G.*; 3D- Segmentation and Vektorvalued Scoring Funktions for Symbolic Docking of Proteins*, GBF Monographs Volume 18, 1995, pp. 105-124.

[Con 86] Connolly M. L.: *Measurement of protein surface shape by solid angles*, Journal of Molecular Graphics, Vol. 4, No. 1, 1986, pp. 3-6.

[EKSX 95] Ester M., Kriegel H.-P., Seidl T., Xu X.: *Feature Based Search for Complementary 3D Surfaces in a Protein Database*, (in German), Proc. Databases in Office, Technics and Science, Springer Verlag, 1995, pp. 373-382.

[KSS 95] Kriegel H.-P., Schmidt T., Seidl T.: *Molecular Surfaces in an Object Oriented Protein Database*, (in German), Proc. Bioinformatics 95, Braunschweig, 1995.

[Koe 90] Koenderink J. J.: *Solid Shape,* MIT Press, Cambridge, MA, 1990.

[MS 95] Meyer M., Schomburg D.: *Protein Docking with Correlation Methods*, Proc. Bioinformatics 95, Braunschweig, 1995.

[Qui 85] Quinlan J.R.: *Induction of Decision Trees*, Machine learning 1, 1986, pp. 81 - 106.

[Wir 95] Wirth S.A.: *Algorithms for Classification in Databases and Their Application to a Protein Database,* (in German), Master's Thesis, Institute for Computer Science, University of Munich, 1995.

Some major improvements of ID3 were introduced [Wir 95] for application on large databases:

- ID3 discovers only rules with a confidence of 1.0. However, from biodatabases such rules can hardly be extracted, and rules with lower confidence may be rather interesting. Therefore, we allow to search for rules with any *minimum confidence* which has to be specified by the user.
- In ID3, the condition of a rule is a conjunction of atomic conditions. We allow disjunctions of "neighbouring" attribute values in a condition in order to increase the support of a rule if needed. The user defines the *minimum support* required for all rules.
- ID3 explores the decision tree in a depth-first manner resulting in a large number of passes through the database. Since these passes are expensive for large databases, our algorithm uses a breadth-first search minimzing the number of database passes.

Experiments

We have performed classification experiments using the BIOWEPRO protein database [KSS 95]. Feature values have been clustered into five intervals to yield symbolic attribute values. To obtain a training set, 22 representative proteins with known docking sites were selected. For each surface point of these proteins, an additional attribute had to be computed specifying whether it belongs to a docking site or not. The resulting training set includes 111000 suface points, 37000 of them belonging to a docking site.

In the following, we present one of the interesting classification rules discovered as an example and show how this rule is extended when a higher confidence is required:

$$SA\_6 = \text{very-high} \wedge HY\_PHOB = \text{very-high} \rightarrow DOCKING = \text{no (with } s = 6124,\ c = 0.86)$$

The price for a higher confidence is a lower support. If, e.g., a minimum confidence of 0.9 is specified then the algorithm chooses SA_12 as the next attribute and extends the rule, e.g., by the condition 'SA_12 = very-high'. Then, the confidence c increases to 0.92, but the support s decreases to 4663. The next expansion would choose 'ELPOT = positive', yielding values of c = 0.95 and s = 2389. Note that the specification of the minimum support is rather crucial because a rule with a very small support usually has a relatively high confidence. However, such a rule will be only applicable to very few objects and, therefore, is not quite useful.

Several rules of very high confidence (> 0.95) were discovered which can be used to discard a significant number of surface segments from the feature index reducing the run time of the filter step of query processing.

Rules of lower confidence cannot directly be used for query processing but present interesting insights into the properties of docking and non-docking sites. The rules discovered indicate that very

- The features values of a surface point depend on the *local neighborhood* of that point.
- Relevant features may be *geometric, physical* and *chemical* attributes of the protein.
- For each feature value there exists a *unique complementary* value in the allowed domain.

For our problem, we have investigated the following features: Solid angle (SA) [Con 86] with different radii for the neighborhood (in Angström), the Local Shape Index (LSI) [Koe 90] with different radii for the neighborhood (in Angström), the Electrostatic Potential (ELPOT) and the Hydrophobicity (HY_PHOB). The choice of different radii for the geometric features was motivated by the observation of a smoothing effect when considering more and more neighbors. Thus, different radii will usually result in different feature values according to the locality of the properties.

Classification

A classification algorithm assumes a database of objects that are described by a collection of attributes each having a small domain of discrete values. Each object belongs to one of a set of mutually exclusive *classes*. The *training set* is a subset of objects of the database whose class is known. The task is to derive a set of classification rules that can determine the class of any object from the values of its attributes. In BIOWEPRO, e.g., the attributes of the surface points are their features and all points may be partitioned into two classes: 1. points at docking sites, 2. points not at docking sites. Then, a classification rule allows to predict whether a point not contained in the training set belongs to a docking site or not.

The basis for our algorithm is *ID3*, a top-down classifier [Qui 85]. ID3 builds a decision tree from the training set. Nodes of the decision tree represent attributes and for each value of this attribute an arc is constructed. Starting from an empty tree, ID3 chooses the attribute with the largest information gain and extends the tree appropriately. For each of the leaves of the current tree, the same procedure is applied considering only the elements of the training set having the respective attribute value. This recursive process continues until all nodes represent members of only one class, i.e. until the whole training set is correctly classified by the descision tree.

From a constructed decision tree as many rules as leaves can be derived by following all pathes from the root to a leaf. A *classification rule* is of the form

$$\text{atrribute}_1 = \text{value}_1 \wedge \ldots \wedge \text{atrribute}_n = \text{value}_n \rightarrow \text{CLASS} = \text{class}_m \text{ (support = s, confidence = c)}$$

The *support s* of a rule r is defined as the relative number of objects in the training set which fulfill the left side (*condition*) and the right side (*conclusion*) of r. The *support for condition* is the relative number of objects in the training set which fulfil the condition and the *confidence c* of r is defined as *support (r) / support for condition (r).*

Published in Proceedings of German Conference on Bioinformatics (GCB-96)

FEATURE BASED CLASSIFICATION OF PROTEIN DOCKING SITES: AN ALGORITHM FOR LARGE DATABASES AND EXPERIMENTAL RESULTS

Martin Ester, Hans-Peter Kriegel, Stefan Wirth

Institute for Computer Science, University of Munich, Oettingenstr. 67, D-80538 München

{ester | kriegel | wirth}@informatik.uni-muenchen.de

Abstract

We report on the development of an efficient classification algorithm for large databases. The algorithm has been used to classify protein docking sites based on a set of geometric and physico-chemical surface features. In our experiments we discovered several classification rules which (1) provide interesting insights into the nature of docking sites and (2) can be used to speed up query processing in a protein docking database.

Introduction

At a biochemical level, many processes of life are based on protein-protein interactions. Therefore, the protein docking problem is attracting more and more interest in the bioinformatics community. The goal of the Project BIOWEPRO is the development of a protein docking system, i.e. for a given query protein all possible docking candidates are extracted from a large database of proteins. To model proteins for the purpose of docking, an appropriate representation of properties of the surface has to be used. In BIOWEPRO, the query processing is divided into three major steps ([AHPS 95], [EKSX 95], [MS 95]):

1. A cheap *filter step* reduces the enormous number of proteins to a moderate number of candidates guaranteeing no false dismissals.
2. An expensive but *more exact test* of the geometry is only performed on those candidates.
3. The *energetically best complex* has to be calculated for the remaining pairs of proteins.

When following a feature based approach, some important questions arise: which features of the surface are suitable to distinguish docking segments from non-docking segments (selectivity), and which of the features correlate and thus are redundant (efficiency)? We have developed an efficient classification algorithm for large databases which was used to find answers to these questions.

Features

Features to support query processing in a protein docking database must fulfill several requirements:

- The feature values are *invariant* with respect to translation and rotation.

convex (SA_6 = very-low) or strongly hydrophob (HY_PHOB = very-high) segments of the surface are likely to be docking sites. On the other hand, if a surface segment is very concave (SA_6 = very-high) or plane and not strongly hydrophob (SA_6 = medium $\wedge$ HY_PHOB = low) or convex and not strongly hydrophob (SA_6 = low $\wedge$ HY_PHOB = low) then it usually is not a docking site. Thus, different types of docking and non-docking sites were discovererd.

The classification algorithm always chose SA_6 as the first attribute and typically chose HY_PHOB as the second one. We conclude that SA_6 and HY_PHOB include the most information (w.r.t. the features investigated) for the classification of docking sites.

Conclusions

We have developed an efficient classification algorithm for large databases. The algorithm was used to classify protein docking sites based on a set of geometric and physico-chemical surface features. Although only a small set of features was investigated, our experiments yielded some interesting and useful classification rules. More features should be considered to increase the confidence of the rules. The following features seem to be potentially useful: carbon, oxygen, and sulphur content, hydrogen-ion-donor or acceptor property and aminoacid content in a defined neighborhood of a surface point.

References

[AHPS 95] Ackermann F., Herrmann G., Posch S., Sagerer G.*; 3D- Segmentation and Vektorvalued Scoring Funktions for Symbolic Docking of Proteins*, GBF Monographs Volume 18, 1995, pp. 105-124.

[Con 86] Connolly M. L.: *Measurement of protein surface shape by solid angles*, Journal of Molecular Graphics, Vol. 4, No. 1, 1986, pp. 3-6.

[EKSX 95] Ester M., Kriegel H.-P., Seidl T., Xu X.: *Feature Based Search for Complementary 3D Surfaces in a Protein Database*, (in German), Proc. Databases in Office, Technics and Science, Springer Verlag, 1995, pp. 373-382.

[KSS 95] Kriegel H.-P., Schmidt T., Seidl T.: *Molecular Surfaces in an Object Oriented Protein Database*, (in German), Proc. Bioinformatics 95, Braunschweig, 1995.

[Koe 90] Koenderink J. J.: *Solid Shape,* MIT Press, Cambridge, MA, 1990.

[MS 95] Meyer M., Schomburg D.: *Protein Docking with Correlation Methods*, Proc. Bioinformatics 95, Braunschweig, 1995.

[Qui 85] Quinlan J.R.: *Induction of Decision Trees*, Machine learning 1, 1986, pp. 81 - 106.

[Wir 95] Wirth S.A.: *Algorithms for Classification in Databases and Their Application to a Protein Database,* (in German), Master's Thesis, Institute for Computer Science, University of Munich, 1995.

Some major improvements of ID3 were introduced [Wir 95] for application on large databases:

- ID3 discovers only rules with a confidence of 1.0. However, from biodatabases such rules can hardly be extracted, and rules with lower confidence may be rather interesting. Therefore, we allow to search for rules with any *minimum confidence* which has to be specified by the user.
- In ID3, the condition of a rule is a conjunction of atomic conditions. We allow disjunctions of "neighbouring" attribute values in a condition in order to increase the support of a rule if needed. The user defines the *minimum support* required for all rules.
- ID3 explores the decision tree in a depth-first manner resulting in a large number of passes through the database. Since these passes are expensive for large databases, our algorithm uses a breadth-first search minimzing the number of database passes.

Experiments

We have performed classification experiments using the BIOWEPRO protein database [KSS 95]. Feature values have been clustered into five intervals to yield symbolic attribute values. To obtain a training set, 22 representative proteins with known docking sites were selected. For each surface point of these proteins, an additional attribute had to be computed specifying whether it belongs to a docking site or not. The resulting training set includes 111000 suface points, 37000 of them belonging to a docking site.

In the following, we present one of the interesting classification rules discovered as an example and show how this rule is extended when a higher confidence is required:

SA_6 = very-high $\wedge$ HY_PHOB = very-high $\rightarrow$ DOCKING = no (with s = 6124, c = 0.86)

The price for a higher confidence is a lower support. If, e.g., a minimum confidence of 0.9 is specified then the algorithm chooses SA_12 as the next attribute and extends the rule, e.g., by the condition 'SA_12 = very-high'. Then, the confidence c increases to 0.92, but the support s decreases to 4663. The next expansion would choose 'ELPOT = positive', yielding values of c = 0.95 and s = 2389. Note that the specification of the minimum support is rather crucial because a rule with a very small support usually has a relatively high confidence. However, such a rule will be only applicable to very few objects and, therefore, is not quite useful.

Several rules of very high confidence (> 0.95) were discovered which can be used to discard a significant number of surface segments from the feature index reducing the run time of the filter step of query processing.

Rules of lower confidence cannot directly be used for query processing but present interesting insights into the properties of docking and non-docking sites. The rules discovered indicate that very

- The features values of a surface point depend on the *local neighborhood* of that point.
- Relevant features may be *geometric, physical* and *chemical* attributes of the protein.
- For each feature value there exists a *unique complementary* value in the allowed domain.

For our problem, we have investigated the following features: Solid angle (SA) [Con 86] with different radii for the neighborhood (in Angström), the Local Shape Index (LSI) [Koe 90] with different radii for the neighborhood (in Angström), the Electrostatic Potential (ELPOT) and the Hydrophobicity (HY_PHOB). The choice of different radii for the geometric features was motivated by the observation of a smoothing effect when considering more and more neighbors. Thus, different radii will usually result in different feature values according to the locality of the properties.

Classification

A classification algorithm assumes a database of objects that are described by a collection of attributes each having a small domain of discrete values. Each object belongs to one of a set of mutually exclusive *classes*. The *training set* is a subset of objects of the database whose class is known. The task is to derive a set of classification rules that can determine the class of any object from the values of its attributes. In BIOWEPRO, e.g., the attributes of the surface points are their features and all points may be partitioned into two classes: 1. points at docking sites, 2. points not at docking sites. Then, a classification rule allows to predict whether a point not contained in the training set belongs to a docking site or not.

The basis for our algorithm is *ID3*, a top-down classifier [Qui 85]. ID3 builds a decision tree from the training set. Nodes of the decision tree represent attributes and for each value of this attribute an arc is constructed. Starting from an empty tree, ID3 chooses the attribute with the largest information gain and extends the tree appropriately. For each of the leaves of the current tree, the same procedure is applied considering only the elements of the training set having the respective attribute value. This recursive process continues until all nodes represent members of only one class, i.e. until the whole training set is correctly classified by the descision tree.

From a constructed decision tree as many rules as leaves can be derived by following all pathes from the root to a leaf. A *classification rule* is of the form

$$\text{atrribute}_1 = \text{value}_1 \wedge \ldots \wedge \text{atrribute}_n = \text{value}_n \rightarrow \text{CLASS} = \text{class}_m \text{ (support = s, confidence = c)}$$

The *support s* of a rule r is defined as the relative number of objects in the training set which fulfill the left side (*condition*) and the right side (*conclusion*) of r. The *support for condition* is the relative number of objects in the training set which fulfil the condition and the *confidence c* of r is defined as *support (r) / support for condition (r)*.

Published in Proceedings of German Conference on Bioinformatics (GCB-96)

FEATURE BASED CLASSIFICATION OF PROTEIN DOCKING SITES: AN ALGORITHM FOR LARGE DATABASES AND EXPERIMENTAL RESULTS

Martin Ester, Hans-Peter Kriegel, Stefan Wirth

Institute for Computer Science, University of Munich, Oettingenstr. 67, D-80538 München

{ester | kriegel | wirth}@informatik.uni-muenchen.de

Abstract

We report on the development of an efficient classification algorithm for large databases. The algorithm has been used to classify protein docking sites based on a set of geometric and physico-chemical surface features. In our experiments we discovered several classification rules which (1) provide interesting insights into the nature of docking sites and (2) can be used to speed up query processing in a protein docking database.

Introduction

At a biochemical level, many processes of life are based on protein-protein interactions. Therefore, the protein docking problem is attracting more and more interest in the bioinformatics community. The goal of the Project BIOWEPRO is the development of a protein docking system, i.e. for a given query protein all possible docking candidates are extracted from a large database of proteins. To model proteins for the purpose of docking, an appropriate representation of properties of the surface has to be used. In BIOWEPRO, the query processing is divided into three major steps ([AHPS 95], [EKSX 95], [MS 95]):

1. A cheap *filter step* reduces the enormous number of proteins to a moderate number of candidates guaranteeing no false dismissals.
2. An expensive but *more exact test* of the geometry is only performed on those candidates.
3. The *energetically best complex* has to be calculated for the remaining pairs of proteins.

When following a feature based approach, some important questions arise: which features of the surface are suitable to distinguish docking segments from non-docking segments (selectivity), and which of the features correlate and thus are redundant (efficiency)? We have developed an efficient classification algorithm for large databases which was used to find answers to these questions.

Features

Features to support query processing in a protein docking database must fulfill several requirements:

- The feature values are *invariant* with respect to translation and rotation.