論 文 内 容 の 要 旨

博士論文題目　Studies on Prediction of Protein Function Based on Oligopeptides
（オリゴペプチドに基づくタンパク質機能予測に関する研究）

氏　名　蓬莱　尚幸

（論文内容の要旨）

タンパク質の機能予測は、バイオインフォマティクスの重要な研究課題である。広範囲なタンパク質機能に対して、タンパク質がもつさまざまな性質を元にした予測が行われているが、本論文では配列のみからの機能予測を扱う。既存技術である相同性検索は比較的長い部分配列の類似性に基づき、パターンマッチングは明確な意味を持つ比較的短い部分配列の存在性に基づいているのに対して、本論文で提案した予測手法は短い部分配列の複数タンパク質における共起関係の偏りに基づいている。タンパク質をそれに含まれるオリゴペプチドの集合ととらえ、それらのオリゴペプチドはそのタンパク質がその機能をもつことを支持しているとみなす。全タンパク質についてオリゴペプチドに分解し、オリゴペプチドごとに集計することで、あるオリゴペプチドに関して、それを含むタンパク質数とそれを含みかつ機能ももつタンパク質数を得る。この比はオリゴペプチドが自分を含むタンパク質が機能をもつことに関する支持の割合を現し、オリゴペプチドと機能の関連度と呼んだ。あらかじめ既知情報からこの関連度を計算しておく。予測アルゴリズムでは、あるタンパク質に含まれる全オリゴペプチドの関連度の和をオリゴペプチド数で標準化することで、そのタンパク質のスコアを計算する。本手法では、このスコアが高いタンパク質がよりその機能を持つ度合いが高いと予測する。

ヒトの標準的なタンパク質を利用して、さまざまな酵素活性およびGene Ontology用語に対する提案手法による予測結果を計測し評価した。予測精度評価には、情報検索分野で提案され通常的に利用されている手法である精度-再現率グラフと最大 F 値を用いた。これらの予測実験により、提案した手法は多くのタンパク質機能に関して非常に有効であることが示された。

提案した手法の予測性能を客観的に示すために、本論文では、既存の予測手法である相同性検索およびパターンマッチングと比較した。対照的予測実験により、提案した手法はパターンマッチングより非常に有効であり、相同性検索に対しても同等以上の性能を持つことが示唆された。

オリゴペプチドの長さは、提案した手法における重要なパラメータであり、予測性能に影響を与える。長さ1から9のオリゴペプチドについて、予測性能を比較した。長さ 1 のオリゴペプチドには予測能力はなく、長さ5以上ではほぼ同等であることが明示された。また、長さ2から4の予測性能は、機能により異なるが、いずれの場合も長さと予測性能は正の相関を示すことも判明した。また、オリゴペプチドが長くなるにつれ、共有性は低下し、予測手法適用性の低下を招くと考えられる。長さ7以上においてオリゴペプチド共有性が極めて低いことを考慮に加え、任意のタンパク質機能の予測に対して汎用的に利用すべきオリゴペプチドの長さは5または6であることが示唆された。

さらに、オリゴペプチドと機能との関連度を利用して、機能に特有かつ一般的なオリゴペプチドを提示する手法を、例を用いて示した。

（1，200字程度）

氏 名	蓬莱 尚幸

（論文審査結果の要旨）

タンパク質機能予測はバイオインフォマティクスの重要な研究課題であり、さまざまな性質を元にした予測が行われているが、本論文では配列のみからの機能予測を扱っている。本論文で提案した予測手法は、あらゆる短い部分配列の複数タンパク質における共起関係の偏りに基づいている。タンパク質をそれに含まれるオリゴペプチドの集合ととらえ、それらのオリゴペプチドはそのタンパク質がその機能をもつことを支持しているとみなす。全タンパク質についてオリゴペプチドに分解し、オリゴペプチドごとに集計することで、あるオリゴペプチドに関して、それを含むタンパク質数とそれを含みかつ機能ももつタンパク質数を得る。この比はオリゴペプチドが自分を含むタンパク質が機能を持つことに関する支持の割合を現しており、オリゴペプチドと機能の関連度と呼んでいる。あらかじめ既知情報からこの関連度を計算しておく。予測アルゴリズムでは、あるタンパク質に含まれる全オリゴペプチドの関連度の和をオリゴペプチド数で標準化することで、そのタンパク質のスコアを計算する。

あらゆる短い部分配列の複数タンパク質における共起関係の偏りに基づく本手法は、既存技術である比較的長い部分配列の類似性に基づく相同性検索や、明確な意味をもつ比較的短い部分配列の存在性に基づくパターンマッチングと、異なる発想の高いオリジナリティーがある。

本論では、ヒトの標準的なタンパク質を利用して、さまざまな酵素活性および Gene Ontology 用語に対する提案手法による予測結果を計測し評価している。予測精度評価には、情報検索分野で提案され通常的に利用されている手法である精度-再現率グラフと最大 F 値を用いており、これらの予測実験により、提案した手法は多くのタンパク質機能に関して非常に有効であることが示されている。

提案した手法の予測性能を客観的に示すために、本論文では、既存の予測手法である相同性検索およびパターンマッチングと比較検討されている。その結果、対照的予測実験により、提案した手法はパターンマッチングより非常に有効であり、相同性検索に対しても同等以上の性能をもつことが示唆された。

オリゴペプチドの長さについても検討されており、長さ 1 のオリゴペプチドには予測能力はなく、長さ5以上ではほぼ同等であることが明示された。また、長さ2から4の予測性能は、機能により異なるが、いずれの場合も長さと予測性能は正の相関を示すことも示された。また、オリゴペプチドが長くなるにつれ、共有性は低下し、予測手法適用性の低下を招くと考えられる。長さ7以上においてオリゴペプチド共有性が極めて低いことを考慮に加え、任意のタンパク質機能の予測に対して汎用的に利用すべきオリゴペプチドの長さは5または6であることが示唆された。

さらに、オリゴペプチドと機能との関連度を利用して、機能に特有かつ一般的なオリゴペプチドを提示する手法を例を用いて示した。

このように、本論文はタンパク質機能予測の情報科学的な推定法の開発において独創性が高いだけでなく、従来研究に対して十分に性能面について優れている。さらにその性能向上の理由についての考察も行った。よって、本論文は博士（理学）の学位論文として価値のあるものと認める。

（A4 1枚 1，200字程度）