質疑応答

Q. タンパク質の４次構造予測では、予測が概ね一致していますが、少しずれがあります。計算では、エネルギーが最小化する構造が出力されるのではないかと思いますが、これは、実際の構造がエネルギー最小ではない構造をとっているということなのでしょうか？

Ans. ここでは計算物理学的なシミュレーションでイメージされるようなエネルギー関数を用いているのではなく，何らかの経験的なエネルギー関数や構造評価関数のようなものを用いています．実際の構造が，（評価関数の設計の問題などで）評価関数が最小となるような構造をとっていないという場合も多々あります．予測構造を起点に，計算物理学的なシミュレーションを行うことで，より実際の構造に近くなるかもしれません．

Q. テンプレートを使って３量体の構造をうまく予測できたとのことですが、そのテンプレートを使うとうまくいくということは、どのようにしてわかるのでしょうか？

Ans.配列検索の結果に基づいています．特にここでは，既知構造の中のどの構造をテンプレート構造として使うかは，既知構造の配列あるいは構造から作られるプロファイルと，予測しようとしている配列のプロファイルとを用いた，プロファイルプロファイルアラインメントを行っています．この結果，アラインメントが有意に良い既知構造は，テンプレートとして有用な可能性があると考え，そのアラインメントと既知構造に基づいて構造予測を行っています．

Q. ディープラーニングによる判断では、プログラムが出してきた結果が、なぜその結果が出てきたのか、理由を知ることができないと聞きますが、今回の例でもそれは同じでしょうか？素人感覚では、なぜそのようにアノテーションされたのかが分からないと、やや気持ちが悪い感じがするのですが、それは仕方ないのでしょうか？

Ans. 学習したモデルが，ある入力配列のどこを重視してどのように重みを付けたことによって結果のアノテーションが出力されたかは，結果を解析することである程度は分かると考えています．一方で，なぜそのような重みが付くようなモデルが，学習データから学習されたかについての理由を知ることは非常に難しいと考えています．

Q. UniProtのTrEMBLのグラフにあるピークはなんですか？

Ans. ピークが表れている2015年4月から，冗長性の除き方に関して新たな考え方が導入されました．具体的には，プロテオーム間での冗長性を定義しており，これに基づいて，細菌，古細菌，真菌では同じ種の異なる株に由来しており，かつ冗長だと考えらえる配列は除く方針になりました．詳しくはUniProtのこちらのページをご覧ください．
https://www.uniprot.org/help/proteome_redundancy

Q. 隠れマルコフモデルというものが出てきますが、「隠れ」の意味は何ですか？

Ans. マルコフ連鎖ではある状態と出力が1対1対応しているのですが，隠れマルコフモデルでは隠れ状態と呼ばれる状態と出力が必ずしも1対1対応してはいません．出力だけを見たときにどの状態から出力されたのか分からないため，隠れ状態と呼ばれます．今回紹介したような配列アラインメントにおける隠れマルコフモデルでは，一般的には，配列のアラインメントされたカラムごとに，一致状態，挿入状態，欠失状態といった隠れ状態が想定され，それらから出力確率に従って確率的に配列/アラインメントが出力されるものとしてアラインメントされた配列をモデル化しています．

►【北日本支部】2020年度オンラインシンポジウム「情報科学を駆使して生命分子を見る・知る・使う」