平成 24 年度　京都大学化学研究所 スーパーコンピュータシステム 利用報告書

新規オーソログクラスタデータベース KEGG OC の構築
Construction of KEGG OC; a novel ortholog cluster database

化学研究所付属バイオインフォマティクスセンター化学生命科学　守屋 勇樹

概要

共通祖先では一つであった遺伝子が種分岐により複数の種に存在している場合、これらの相同配列をオーソログ遺伝子と呼ぶ。新規に読まれた配列におけるオーソログの同定は、その機能解析に重要である。また正しいオーソログを推定することは配列や配列の作り出すネットワークにおける比較解析でも重要な要素である。そのためこれまで COG や eggNOG といった多くのオーソログクラスタデータベースが作成されている。しかしこれらのデータベースは扱っている種数や、手作業の必要性による更新の遅れなどの問題を抱えている。そのため、これらの問題を回避した新たなオーソログクラスターデータベース、KEGG OC を作成した。KEGG OC は KEGG に登録されているコンプリートゲノム全てを用いることで、これまでで最大のデータベースとなっている他、パスウェイや遺伝子の階層分類などの KEGG の他のデータベースを利用できるようになった。また KEGG データベースに手作業を加えることなく、完全な自動計算によって構築することが可能であり、これにより定期的な更新が可能になった。

手法

KEGG OC は系統樹に従って分類群毎にクラスタリングを行うことで、コストをかけることなく計算を行なっている。図 1 は用いた系統樹の例で、動物の場合まず種内の配列でクラスタリング行い、次に鋼の階級で種の階級で作られたクラスタのクラスタリングをこなうことで、一段階上位の分類群クラスタを作成した。クラスタリングを門、界、ドメインの階級まで繰り返すことで最終的なオーソログクラスタを得た。各分類郡におけるクラスタリングには Quasi-clique-based クラスタリングを用い、配列間の Smith-Waterman スコアからなる類似度ネットワークから準クリークを抽出することで行った[1]。また、多数の下位分類群クラスタを一度にクラスタリングする場合など、類似度ネットワークは極端に巨大になるため、Inparanoid アルゴリズムによりエッジの削減を行った[1]。

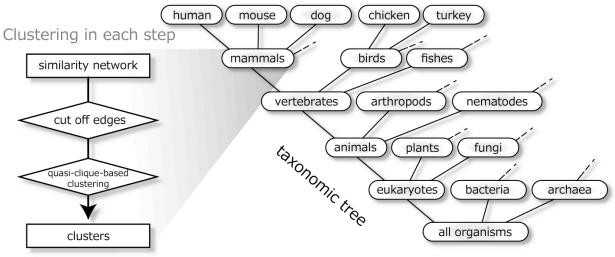

図 1 ：KEGG OC 構築の流れ

結果と考察

約 900 万のタンパク質配列から 124 万のオーソログクラスタを作成できた。表 1 は作成されたクラスタの例で多くのアクアポリンタンパク質から成っている。一つのクラスタは一つまたは複数の下位の分類群クラスタから成っており、これら一つのクラスタにまとめられたアクアポリンタンパク質が共通祖先において一つの配列であったことを示唆している。また、動物分岐以後、脊椎動物分岐以前に配列が 2 つに分岐し、されにその内片方が哺乳類分岐以前に 3 つに分岐したことも示唆している。このように KEGG OC は全てのタンパク質ファミリにおいて配列の進化イベントの情報

OC	TC 1	TC 2	TC 3	TC 4	PC	gene
OC. 219800	Eukaryotes. 77375	Animals.41203	Vertebrates.2493	Mammals. 26632	hsa.8548	hsa:359
					ptr.3412	ptr:451886
				Mammals. 17805	hsa.8585	hsa:362
					ptr.17687	ptr:741283
				Mammals. 18101	hsa.8595	hsa:363
					ptr.17711	ptr:741338
			Vertebrates.5483	Mammals. 9399	hsa.8577	hsa:361
					ptr.5381	ptr:455350
		⋮	⋮	⋮	⋮	⋮
		Animals.41207	Cnidarians.16648	-	nve.6083	nve:NE...

表１：オーソログクラスタの例

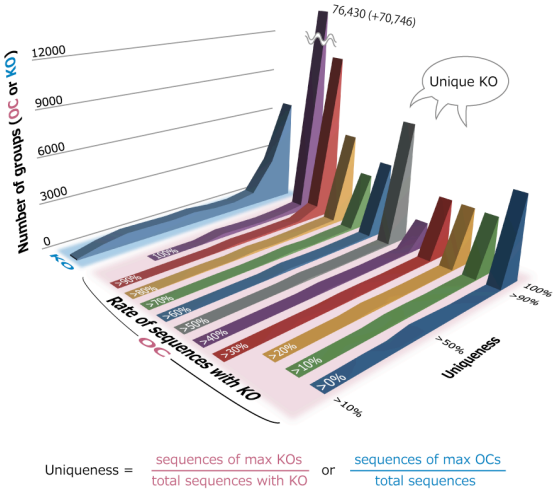

図２：KO との比較

が反映されているといえる。次に手作業で作成された KEGG Orthology (KO)との比較を行った。KO のアサインされたタンパク質を含むクラスタは約 20 万であり、KO の 15,600 グループと比較して細分化されていた。図２は各クラスタに含まれる KO の種類の特異性を示したもので、KO のついている割合に関係なく、ほとんどのクラスタが一種類の KO を持っていた。また逆の KO グループにおけるクラスタの特異性においても、一種類のクラスタを含む KO の割合が多かった。このことから自動計算によって作成されたオーソログクラスターが手作業での分類によって作成された KO のをある程度再現できているといえ、精度の高いオーソログクラスタが構築されていると考えられる。

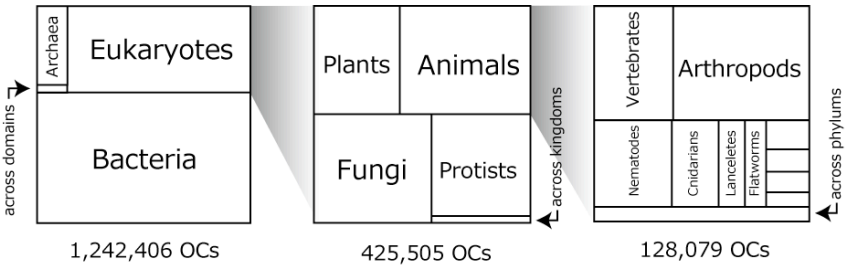

図３：分類群特異的クラスタの分布

図３は分類群特異的クラスタの分布を示しており、各領域の面積がクラスタの数を表している。多くのクラスタが分類群特異的に構築されており、複数のドメイン、界、門を含むクラスタが比較的少ないことが示された。特にドメインを跨ぐクラスタについてはこれまでの知識と比較しても非常に少なくなっており、クラスタリング手法を含め再度検討が必要である。

KEGG OC は Web ベースで提供されており、GenomeNet（http://www.genome.jp/tools/oc/）で利用可能になっている。

発表論文

[1] Nakaya A, Katayama T, Itoh M, Hiranuka K, Kawashima S, Moriya Y, Okuda S, Tanaka M, Tokimatsu T, Yamanishi Y, Yoshizawa AC, Kanehisa M, Goto S. ; KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Res.* (2013).