JP38 生化学・生物学の専門用語と意味関係によって構築したネットワークの変化

(筑波大図書館情報‡、ATR 人間情報科学研†)
○真栄城哲也†‡*、下原勝憲†、中山伸一‡

1. 緒言

生化学および生物学の概念をノードで表し，概念間に意味関係があればリンクで結ぶことによって得られる生化学・生物学の概念ネットワークの特性を解析した．生化学・生物学概念ネットワークの特性は，生化学および生物学の知識全体の特性を概念と意味関係の側面から捉えたものであり，このような特性が明らかになれば，ケモインフォマティクスやバイオインフォマティクスの分野において意味に関する情報処理，表現，記録の最適な手法の手掛かりが得られる．

ここでは，細胞分子生物学という特定の分野において，概念ネットワークが時間とともにどのように変化しているかの解析について報告する．

2. 概念ネットワークの解析

対象としている生化学・生物学のような自然科学の場合には，専門用語が対象分野の概念を表していると考えられる．概念間の意味関係に基づいて概念を構造化して初めて，意味処理に関する情報処理に有効な解析が可能となる．既に，分子生物学，脳神経学，生化学，免疫学，微生物学等の分野を統合した概念ネットワークを構築済である [1]．この概念ネットワークには，133,251 個の概念が 2,574,099 個の意味関係で関連付けられている．抽出した概念間の意味関係の種類は，同義関係，関連関係，類似関係，階層関係，属性，である．なお，概念当りの意味関係の平均数 (ノード当りの平均リンク数) は 19.3 個である．

意味関係数 k から捉えた概念ネットワークの構造が巾分布構造を持ち，スケールフリー構造であることは，概念ネットワークの形成および成長がランダムではないことを示唆している．ネットワークの規模が新しいノードの追加によって拡大していくネットワーク成長過程について計算機シミュレーション実験を行い，新しいノードから既存のノードへの接続と，既存のノード間の新しい接続の計 2 種類の接続を追加する複数のアルゴリズムを比較した結果，少なくともランダムな接続では巾分布構造を持つネットワークは生じないことが判っている．この結果が示唆するのは，概念ネットワークにおいては，新しい概念は意味関係を多く持つハブ概念に選択的に接続しやすいということである．従って，新しい概念は重要な概念に関連付けられて概念ネットワークに組み込まれる．

概念ネットワークにおける概念間の接続特性の解析として，概念がいくつの意味関係を持つ概念と接続しているかを解析したところ，概念ネットワークの成長に関しての洞察が得られた (図 1)．この結果を，今回の対象である細胞分子生物学において検証した．

本研究では，この概念ネットワークの枠組みで，細胞分子生物学の概念ネットワークがどのように変化してきたかの解析をする．3 段階で計測するデータは，概念数でみると，それぞれ 9,727 個，10,983 個，14,373 個と増加している．構築済の概念ネットワーク内での変化を分析することで，細胞分子生物学よりも広い分野である生物学の中での挙動を知ることができる．

解析は，ノード間の接続に注目した手法を用いる．1 つは，ノード当りのリンク数の分布であり，これは概念がどれだけ密接に他の概念と意味的に関連しているかの分布である．接続が密な概念ほど，重要な概念であると捉えている．もう 1 つの手法は，ある概念が，同数以上のリンク数を持つ概念と接続している割合である．この解析は，ノード当りのリンク数がその概念の重要性を反映しているとの考えに基づいている．また，リンク数の分布が巾分布になるには，新しいノードがネットワークに導入される際，接続先のノードはリンク数に比例して選択されるためである．

3. 考察および結論

概念毎の意味関係の数 k のヒストグラムを取ると，1,000 個前後の意味関係を持つ概念を境に 2 重のスケールフリー構造を持つことを既に報告しており，概念ネットワークは Small World

*maeshiro@slis.tsukuba.ac.jp

Network (SWN) [2,3] と類似の構造を持っている [5].

このことは，概念間の距離を，概念を接続する意味関係の最短数と定義した場合，多くの概念と意味関係を持つハブとして機能する少数の概念の存在によって，ある概念から他の概念へ短距離で到達できることを意味する．これらのハブ概念は，生化学・生物学の知識体系で重要な概念であり，生化学・生物学の専門家ならば必ず知らなければならない概念である．これは，SWN の特徴であるハブノードの構造である．また，本研究で抽出した意味関係を用いた場合，生物学の重要な概念は他の重要な概念を直接連想させると考えられる．このことは，特定の概念の重要性がネットワーク構造から推測できることを示唆している．本手法は，このように意味関係の数によって概念の重要性を定量化する手法である．

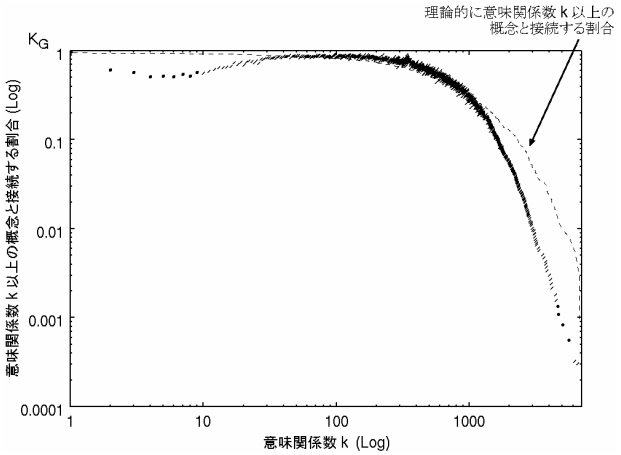

Figure 1. 概念の接続と意味関係数の関係. ある概念と接続している概念について，意味関係数 k より多い意味関係を持つ割合 (KG). 点線は，概念ネットワーク全体で意味関係数 k より多い意味関係を持つ概念の割合を理論的に計算したもの

意味関係数 k が与えられた時，接続している概念の意味関係数が，対象としている概念の持つ意味関係数 k より多いか少ないかで分類し，対象の概念と同じクラス (k 以下) に接続している割合を計算する．図 1 は，意味関係数を k 個持つ概念に接続している k 個の概念の内，k 個より多い意味関係数を持つ概念の割合である．

意味関係数 2,000 個前後以上のハブノードの領域に巾分布構造が見られる．また，ランダム性を意味するポアソン分布 [4] とは明らかに異なる．

一方，意味関係数が少ない概念においても，構造に特性が見られる．図 1 の点線は概念ネットワーク全体で意味関係数 k より多い意味関係を持つ概念の割合を重み付きで計算した値であり，この比較の理由は，SWN の形成方法として概念間 (ノード間) の接続確率が意味関係数 (リンク数) に比例する接続形態が知られているからである．

概念ネットワークの形成が，SWN の 1 形成方法である接続確率が意味関係数 (リンク数) に比例した接続によるならば両曲線は同じはずであるが，図 1 の実線は単純減少せずに，k=10 以降で一旦増加する．意味関係数を 1 つしか持たない概念の殆んどが意味関係を 2 つ以上持つ概念と接続しているのは当然である．しかし，意味関係数 k が 2 から 15，特に $2 \leq k \leq 10$ の範囲で同じか少ない意味関係数を持つ概念と継っている割合が今回構築した概念ネットワークの特徴である．従って，$2 \leq k \leq 15$ の範囲の意味関係数を持つ概念は小規模なクラスタを形成しており，k=20　前後を境にした階層構造を持つと考えられる．

意味関係数が 1 つの概念には、物質名が多く，特に用語を構成する単語数が多い用語が確認される．物質名に関しては，新しい物質は既存の物質からの派生の形で命名されることが多い．新しい物質が発見されると，意味関係数 1 の概念として概念ネットワークに関連付けられ，その物質の有効性や性質等の解析によって，意味関係数を増加させていくと推測される．同時に，意味的な関連付けは同じ対象の概念集合内で行われる．この集合をクラスタとすると，1 つのクラスタを形成する概念の規模は，意味関係数 k=15～20 であると考えられる．物質名以外の概念についても，同様だと思われる．

以上の結果から，意味関係数の観点からは，概念が概念ネットワークに組み込まれる過程において，意味関係数 k=15～20 の臨界点の存在が推測される．

今回は，細胞分子生物学の分野において，これらの推測を確認した．

参考文献

[1] 真栄城　他　「情報知識学会誌」13, 1-9 (2003).

[2] D. J. Watts, S. J. Strogatz Nature 393, 440-42 (1998).

[3] R. Albert, A.-L. Barabasi Reviews of Modern Physics 74, 47-97 (2002).

[4] B. Ballabas Random Graphs, Academic Press, (1985).

[5] 真栄城　他　「生化学・生物学の専門用語と意味関係によって構築したネットワークの特性」第 26 回情報化学討論会, 107-108 (2003).