# Adaptation of local clustering techniques to biomedical data sets

## １. 代表研究者名

［国立情報学研究所］HOULE, Michael

## ２. 共同研究者

［国立情報学研究所］安達淳、佐藤真一

［国立遺伝学研究所］菅原秀明、阿部貴志

## ３. 平成 17 年度の研究実績報告

### 3-1.研究目標

The goal of this project was to extend a generic clustering technique that Houle had previously developed, “local clustering”, for use on protein sequence data. Local clustering uses neighborhood information to determine local regions of high mutual association among items. Fast, accurate approximations of the neighborhoods can be used to greatly speed the clustering process, and to scale to data of extremely high dimension.

The project involved the creation of clustering tools for protein & gene sequence data, and their implementation as NIG web services. In order to cope with large data sets (millions of proteins), the tool would need to be able to make use of parallelism and external memory.

### 3-2. H17 年度の研究成果（概要）

- The ``local clustering’’ model was further developed into the Relevant Set Correlation (RSC) clustering model, a theoretical framework in which the quality of cluster candidates, the degree of association between pairs of cluster candidates, and the degree of association between clusters and data items are all assessed according to the statistical significance of a form of correlation among neighborhood sets (relevant sets) and/or candidate cluster sets. RSC is much more powerful than the original “local clustering” model, in that it can generate clusters of size as small as 3 or 4 items (the previous lower limit was 25 items).
- A clustering heuristic, GreedyRSC, was developed and implemented based on the RSC model. Although the current implementation is for a single processor, the code as written can be parallelized if needed. The heuristic is implemented so as to make use of external memory, greatly expanding the size of the data sets that can be clustered.
- A DNA Data Bank of Japan web service was developed based on the GreedyRSC clustering tool. The tool allows to search and browse through a static clustering of the entire bacterial ORF database maintained by DDBJ. It also allows users to upload protein sequences for vectorization and clustering by the system, and to browse or download the results. Most of the budget was spent on the development of the web service and on the purchase of a server dedicated to it.

### 3-3. 今後の展開

Planned extensions to this work include:

- Enhancements to the approximate similarity search structure used by GreedyRSC for efficient generation of neighborhood information.
- Adaptation of the clustering heuristic to allow fast updates of clusterings after changes to the data set.
- Development of a new `partitional’ clustering heuristic based on RSC more suited to

applications involving classification of elements.

- Updating the CGM clustering web service to take advantage of enhancements to the underlying clustering tool and model.

### 3-4. 成果発表及び執筆論文

- The CGM clustering web service is accessible at http://rhodem11.ddbj.nig.ac.jp/CGM/.
- The new RSC clustering model and GreedyRSC clustering heuristic is described in: M. E. Houle, "A generic query-based model for scalable clustering", NII Technical Report NII-2006-008E, 19 May 2006. A submission based on this work is planned for the IEEE ICDM 2006 conference (submission deadline 5 July 2006).
- An application of the RSC clustering model and the GreedyRSC heuristic to human face retrieval: D.-D.¥ Le, S. Satoh and M. E. Houle, "Face retrieval in broadcasting news video by fusing temporal and intensity information", in *Proc. 5th Int'l Conf. on Image and Video Retrieval* (CIVR 2006), Tempe AZ, USA, July 2006, to appear.
- Although the work was performed before this project began, the following relevant paper on the SASH approximate similarity search structure appeared in H17: M. E. Houle and J. Sakuma, "Fast similarity search in extremely high-dimensional data sets",. In *Proc. 21st IEEE Int'l Conf. on Data Engineering* (ICDE 2005), Tokyo, Japan, Apr. 2005, pp. 619-630.