140219 External Review of CCS

Activities and Collaborations

Division of Life Sciences: Molecular Evolution Group

T. Hashimoto

Collaborative Fellow of CCS
Faculty of Life and Environmental Sciences

Members of Molecular Evolution Group

Faculty members of CCS

Yuji Inagaki	Assoc. Prof.
Takuro Nakayama	Researcher (Jun. 2013～)

Faculty members from Faculty of Life and Environmental Sciences

Tetsuo Hashimoto	Prof. (Collaborative Fellow)
Goro Tanifuji	Assist. Prof. (Sep. 2013～)
Ryoma Kamikawa	Assist. Prof. (Dec. 2011～Mar. 2013)
Akifumi Tanabe	Researcher/Assist. Prof. (Apr. 2009～Aug. 2011)
Miako Sakaguchi	Researcher (Apr. 2005～Nov. 2008)

Ph. D. Students

Sohta Ishikawa*	Doctoral Program in Biological Sciences, Graduate School of Life and Environmental Sciences
Yuki Nishimura*	

*Both students belong also to Master's program in Computer Science, Graduate School of Systems and Information Engineering

Research Interests of Molecular Evolution Group

Evolutionary bioinformatics:
Methodological studies in molecular phylogeny

Evolutionary biology:
Evolutionary studies on eukaryotic cells and their genomes

Molecular phylogeny: phylogenetic inference from Sequence data

DNA **[Translation]** **protein**

nucleotide (base) sequence → amino acid sequence
(4 bases) (20 amino acids)
{A,T,C,G} {A,R,N,D,C,Q,…,W,Y,V}

α-hemoglobin alignment

site : evolutionarily related

*site with identical amino acid shared by four animals

Inference of a tree
by a single gene (α-hemoglobin)
Data matrix: (4, 141)
4 species, 141 sites , 1 gene

Phylogenomics
with multi-genes concatenated
50～ species, 30,000～ sites
100～ genes huge matrix

Difference in DNA or protein sequences between organisms
→ information on evolutionary process (tree) of the organisms

Maximum likelihood (ML) method: model for substitution process

Model for process of base/amino acid substitutions

- Transition rate matrix Q

: i to j transition rate during infinitesimally short time interval dt

ex.) General Time Reversible (GTR) model for base substitutions → GTR model for amino acid substitutions

$$Q = \begin{bmatrix} — & ag_T & bg_C & cg_G \\ ag_A & — & dg_C & eg_G \\ bg_A & dg_T & — & fg_G \\ cg_A & eg_T & fg_C & — \end{bmatrix}$$

Q = [20 x 20] matrix

$a \sim f$: model parameters

g_A, g_T, g_C, g_G : equilibrium compositions of bases

- Transition probability matrix P

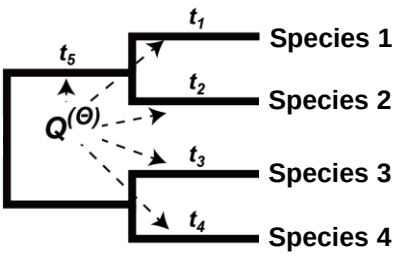
$$P_{ij}(t) = e^{tQ}$$

i, j : states of four bases

t : evolutionary time, number of substitutions/site

Homogeneous and non-homogeneous substitution models

Homogeneous

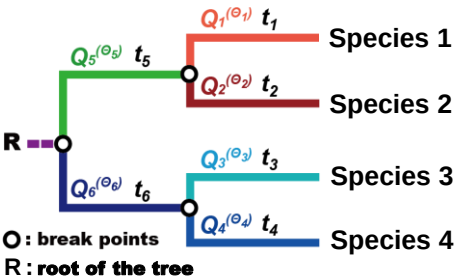


Single matrix $Q^{(\Theta)}$ to all branches

Assumption

Sequences should evolve following same substitution process

Non-Homogeneous



Different matrices $Q_1^{(\Theta_1)} \sim Q_6^{(\Theta_6)}$ to each branch

Assumption

Sequences can evolve following independent processes across tree

Assumptions of the maximum likelihood (ML) method

X_t : state of base/amino acid at time t

[1] X_t is a time-continuous, Markov process with transition probability, $P_{ij}(t)$. Transition from i *to* j is represented by

$$P_{ij}(t) = P\{X_{t+s} = j \mid X_s = i\}.$$

[2] Evolution (substitutions) on each branch occurs independently.

[3] Each site, $X_t^{(h)}$ $(h = 1,\ldots,n)$, evolves independently with an identical probability law.

[4] In the following example, homogeneous model with same $P_{ij}(t)$ across tree is assumed.

ML method: model for branching order of tree (tree topology)

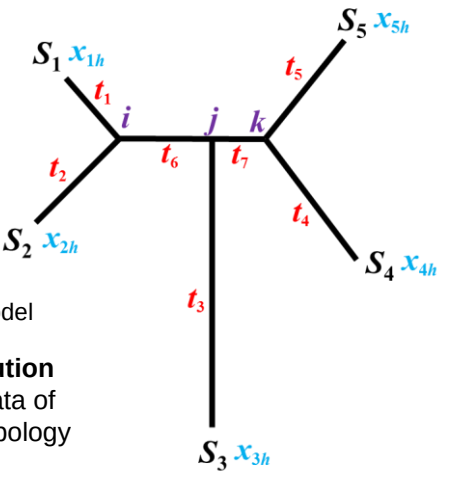
ex.) data matrix (alignment) for 5 species

$\mathbf{X} = (x_{lh})$ $(l = 1,\ldots 5;\ h = 1,\ldots,n)$

- h'th site: $X^{(h)} = (x_{1h}, x_{2h}, x_{3h}, x_{4h}, x_{5h})$

⇒ Data (states) of extant species, $S_1,\ldots, S_5$

- i, j, k : state of h'th site for ancestral species
- $t_1,\ldots, t_7$: branch lengths
- model parameters: θ branch lengths and other parameters related to substitution model



By the assumption of **independent evolution for each branch**, probability of getting data of h'th site with a given $P_{ij}(t)$ and the tree topology shown in right is,

$$f(x_{1h}, x_{2h}, x_{3h}, x_{4h}, x_{5h} \mid \theta)$$

$$= \sum_i \{\pi_i P_{ix_{1h}}(t_1) P_{ix_{2h}}(t_2) \sum_j \{P_{ij}(t_6)\, P_{jx_{3h}}(t_3) \sum_k P_{jk}(t_7)\, P_{kx_{5h}}(t_5) P_{kx_{4h}}(t_4)\}\}$$

(π_i **:composition of base or amino acid** i **)**

ML estimation of parameters

By the assumption of **independent evolution for each site**, probability of getting a data matrix with a given $\boldsymbol{P_{ij}(t)}$ and a given tree topology can be regarded as a function of model parameters:

$$L(\theta \mid \mathrm{X}) = \prod_{h=1}^{n} f(X^{(h)} \mid \theta) \quad : \text{likelihood}$$

$$l(\theta \mid \mathrm{X}) = \sum_{h=1}^{n} \log f(X^{(h)} \mid \theta) \quad : \text{log-likelihood}$$

Estimation of parameters : parameter estimates $\hat{\theta}$ are given by maximizing the log-likelihood function :

$$l(\hat{\theta} \mid \mathrm{X}) = \max_{\theta} l(\theta \mid \mathrm{X})$$

For alternative trees i (i=1,…N) :

N: number of possible tree topologies for a given number of species

$l_i(\hat{\theta}_i \mid \mathrm{X}) \Rightarrow$ compared

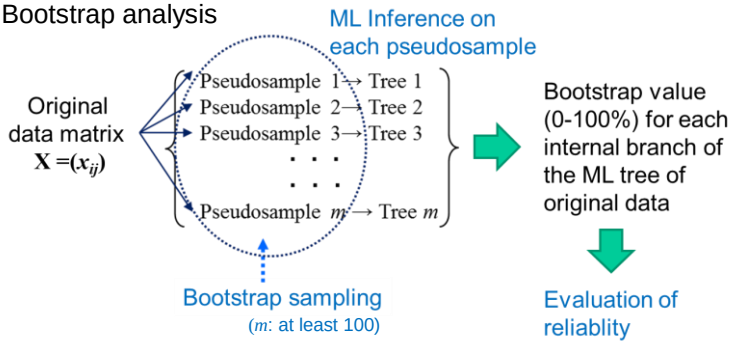
5 species ⇒ 15 trees

$\max_i l_i(\hat{\theta}_i \mid \mathrm{X})$ i (i=1,…N)

⇒ Maximum Likelihood (ML) tree

Phylogenetic analyses are computer intensive

- Huge data matrices in phylogenomics
  (～100 species, ～100,000 sites from ～200 genes)
  Too many tree topologies: exhaustive search impossible
  ⇒ Heuristic tree search (HTS)
- Sophisticate, parameter-rich models
  To avoid misleading inference stemming from model mis-specification
- Reliability of the ML tree
  : Bootstrap analysis



Research activities
1. Methodological studies in molecular phylogeny

- Performance of the ML methods for base sequence data with parallel composition heterogeneity (Ishikawa et al. 2012a,b)
  - Simulation study: non-homogeneous model > homogeneous model
- Parallelization of the NHML program which implements a non-homogeneous base substitution model, GG98 (Ishikawa et al. 2013, 2014)
  - Achievement of the suitable performance of parallelization with more than 1024 cores
- Potential bias in bootstrap support values in the heuristic tree search (HTS)-based ML methods
  - Efficiency of HTS on obtaining correct bootstrap values using simulated datasets
- Dependence of multi-gene phylogeny on gene-sampling (Inagaki et al. 2009)
  - Illustrative data analysis on the issue of archaeplastids monophyly

Publications in methodological studies

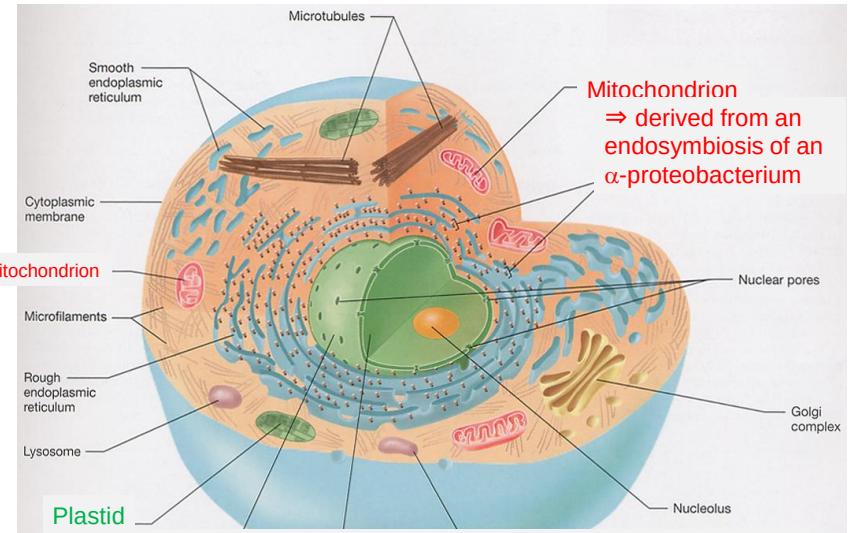
Original papers:

- S. A. Ishikawa et al. MPI/OpenMP HYBRID Parallelization for Phylogenetic Analyses based on Non-Homogeneous Substitution Models:Implementation and Performance Evaluation for Large-Scale Computing Systems. **IPSJ Transactions on Advanced Computing System**, 47, in press (2014)
- S. A. Ishikawa, Y. Inagaki & T. Hashimoto. RY-coding and non-homogeneous models ameliorate the maximum-likelihood inferences from nucleotide sequence data with parallel compositional heterogeneity. **Evolutionary Bioinformatics**, 8, 357-371 (2012)
- Y. Inagaki, Y. Nakajima, M. Sato, M. Sakaguchi & T. Hashimoto. Gene sampling can bias multi-gene phylogenetic inferences: the relationship between red algae and green plants as the case study. **Molecular Biology and Evolution**, 26, 1171-1178 (2009)

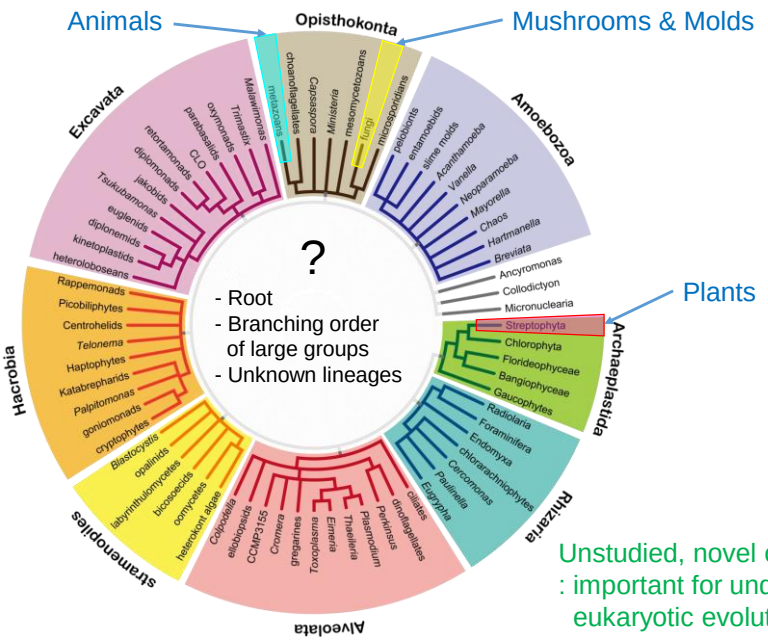
Proceedings:

- S. A. Ishikawa, H. Nakao, Y. Inagaki, T. Hashimoto & M. Sato. Hybrid MPI/OpenMP parallelization of a phylogenetic program with Non-Homogeneous models: toward the analyses of large-scale sequence datasets. **Proceedings of High Performance Computing Symposium** (2014)
- M. Tsuji, M. Sato, A. S. Tanabe, Y. Inagaki & T. Hashimoto. An asynchronous parallel genetic algorithm for the maximum likelihood phylogenetic tree search. **Proceedings of 2012 IEEE Congress on Evolutionary Computation** (2012)
- S. A. Ishikawa, T. Hashimoto. Assessment of the performance of phylogenetic inference based on simulated protein-coding sequences with significant compositional heterogeneity. **Proceedings of the Institute of Statistical Mathematics**, 60, 289-303 (2012) (in Japanese)

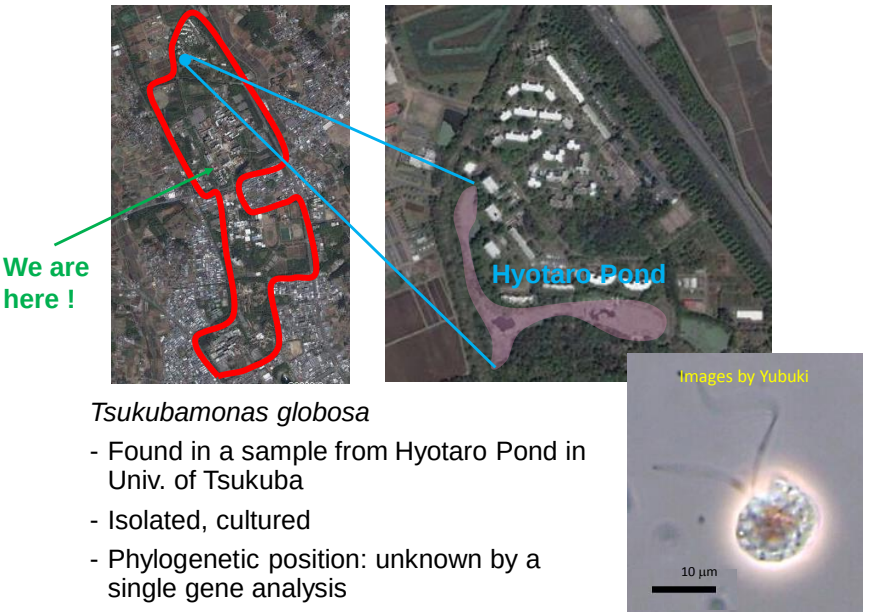
Eukaryotic cell



Diversity of eukaryotes: recent version of eukaryotic tree



Tsukubamonas: a novel eukaryotic microorganism



Tsukubamonas globosa

- Found in a sample from Hyotaro Pond in Univ. of Tsukuba
- Isolated, cultured
- Phylogenetic position: unknown by a single gene analysis

An example of phylogenomic analysis: *Tsukubamonas* ⇒ Discoba

Tsukubamonas globosa

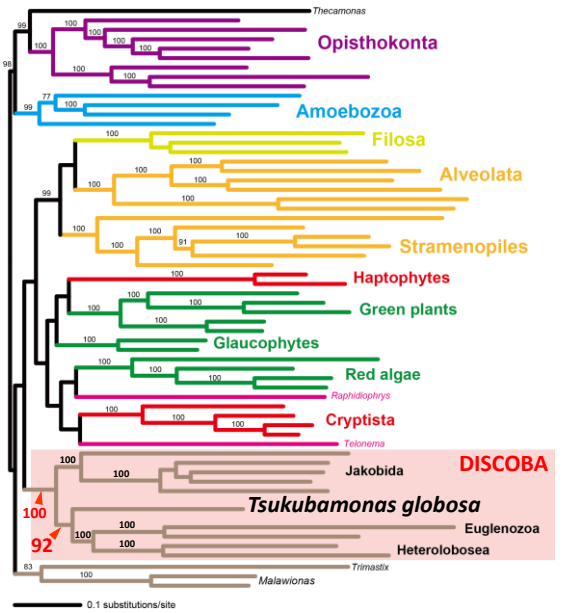
- Isolation of DNA/RNA
- Next generation sequencing (NGS)

⇒ a large amount of sequence data

236,871 reads

12,694 contigs

⇒ a large matrix,
59 species, 41,372 aa sites



ML method, LG + Γ + F model

- Recovered Discoba monophyly
- *Tg* is a novel lineage in Discoba

2. Evolutionary studies on eukaryotic cells and their genomes

- Global eukaryotic phylogeny (Yabuki et al. 2014, 2011, 2010; Kamikawa et al. 2014; Takishita et al. 2012; Burki et al. 2009)

  Phylogenomic analyses including previously unstudied, novel organisms
  - *Palpitomonas bilix* (Hacrobia) [64 species, 41,372 aa sites]
  - *Tsukubamonas globosa* (Discoba) [72 species, 41,372 aa sites]
  - *Carpediemonas*-like organisms (Fornicata) [20 species, 39,089 aa sites]
  - *Raphidiophrys contractilis* (Hacrobia) [75 species, 29,235 aa sites]

- Evolution of mitochondria (Kamikawa et al. 2014; Nishimura et al. 2012;)
  - Comparative transcriptomics analysis of mitochondrion related organelles (MROs) in fornicates
  - Mitochondrial genome sequence analyses of diverse protest lineages and comparative genomics

- Evolution of plastids (Matsumoto et al. 2011; Takishita et al. 2008; Arisue et al. 2012)

  Plastid genome sequence analyses
  - Green algae origin of the dinoflagellate genus *Lepidodinium* plastid
  - Phylogeny of malaria parasites based on the vestigial plastid genome encoded genes

2. Evolutionary studies on eukaryotic cells and their genomes (continued)

- Evolution of bacterial endosymbionts in diverse eukaryotic cells

  ‖

  'Younger organelles' than mitochondria/plastids

  ⇒ Models for early phase of organelle genesis

  ⇒ Recently started new project

  Genome sequence analyses of the cyanobacterial symbionts, which were highly integrated into host (eukaryotic) cells
  - Rhopalodiacean diatoms
  - a testate amoebae, *Paulinella chromatophore*

- Evolution of translation elongation factors in eukaryotes (Kamikawa et al. 2013, 2011, 2008; Sakaguchi et al. 2009)

  Survey of the two elongation factor types, EF-1α and EFL, in diverse eukaryotes
  - Patchy distribution of the two types, EF-1α/EFL, across global eukaryotic tree
  - Independent differential losses of one of the two factors in descendant lineages

Main publications (original papers) in evolutionary studies

- Global eukaryotic phylogeny

  - R. Kamikawa, M. Kolisko, Y. Nishimura, A. Yabuki, M. W. Brown, S. A. Ishikawa, K. Ishida, A. J. Roger, T. Hashimoto & Y. Inagaki. Gene-content evolution in discobid mitochondria deduced from the phylogenetic position and complete mitochondrial genome of *Tsukubamonas globosa*. **Genome Biology and Evolution**, in press (2014)

  - K. Takishita, M. Kolisko, H. Komatsuzaki, A. Yabuki, Y. Inagaki, I. Cepicka, P. Smejkalova, J. D. Silberman, T. Hashimoto, A. J. Roger & A. G. B. Simpson. Multigene phylogenies of diverse *Carpediemonas*-like organisms identify the closest relatives of 'amitochndriate' diplomonads and retortamonads. **Protist**, 163, 344-355 (2012)

  - A. Yabuki, T. Nakayama, N. Yubuki, T. Hashimoto, K. Ishida & Y. Inagaki. *Tsukubamonas globosa* n. g., n. sp., a novel excavate flagellate possibly holding a key for the early evolution in "Discoba." **Journal of Eukaryotic Microbiology**, 58, 319-331 (2011)

  - A. Yabuki, Y. Inagaki & K. Ishida. *Palpitomonas bilix* gen. et sp. nov.: A novel deep-branching heterotroph possibly related to Archaeplastida or Hacrobia. **Protist**, 210, 523-538 (2010)

  - F. Burki, Y. Inagaki, J. Brate, J. M. Archibald, P. J. Keeling, T. Cavalier-Smith, M. Sakaguchi, T. Hashimoto, A. Horak, S. Kumar, D. Klaveness, K. Jakobsen, J. Pawlowski & K. Shalchian-Tabrizi. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, Telonemia and Centroheliozoa, are related to photosynthetic chromalveolates. **Genome Biology and Evolution**, 1, 231-238 (2009)

Main publications (original papers) in evolutionary studies (continued)

- Mitochondrial evolution

  - R. Kamikawa, M. Kolisko, Y. Nishimura, A. Yabuki, M. W. Brown, S. A. Ishikawa, K. Ishida, A. J. Roger, T. Hashimoto & Y. Inagaki. Gene-content evolution in discobid mitochondria deduced from the phylogenetic position and complete mitochondrial genome of *Tsukubamonas globosa*. **Genome Biology and Evolution**, in press (2014)

  - Y. Nishimura, R. Kamikawa, T. Hashimoto & Y. Inagaki. Separate origins of group I introns in two mitochondrial genes of the katablepharid *Leucocryptos marina*. **PLoS ONE**, 7, e37307 (2012)

  - M. Kolisko, J. D. Silberman, I. Cepicka, N. Yubuki, K. Takishita, A. Yabuki, B. S. Leander, I. Inouye, Y. Inagaki, A. J. Roger & A. G. B. Simpson. A wide diversity of previously undetected relatives of diplomonads isolated from marine/saline habitats. **Environmental Microbiology**, 12, 2700-2710 (2010)

- Plastid evolution

  - T. Matsumoto, F. Shinozaki, T. Chikuni, A. Yabuki, K. Takishita, M. Kawachi, T. Nakayama, I. Inouye, T. Hashimoto & Y. Inagaki. Green-colored plastids in the dinoflagellate genus *Lepidodinium* are of core chlorophyte origin. **Protist**, 162, 268-276 (2011)

  - N. Arisue, T. Hashimoto, M. Mitsui, M.L.Q Palacpac, A. Kaneko, S. Kawai, M. Hasegawa, K. Tanabe & T. Horii. Split introns in the genome of *Giarida intestinalis* are excised by spliceosome-mediated trans-splicing. **Molecular Biology and Evolution**, 29, 2095-2099 (2012)

Main publications (original papers) in evolutionary studies (continued)

- EF-1a/EFL evolution
  - R. Kamikawa, M. W. Brown, Y. Nishimura, Y. Sako, A. A. Heiss, N. Yubuki, R. Gawryluk, A. G. B. Simpson, A. J. Roger, T. Hashimoto & Y. Inagaki. Parallel re-modeling of EF-1α function in eukaryotic evolution: Divergent, low-expressed EF-1α genes co-occur with EFL genes in diverse distantly related eukaryotes. **BMC Evolutionary Biology**, 13, 131 (2013)
  - R. Kamikawa, A. Yabuki, T. Nakayama, K. Ishida, T. Hashimoto & Y. Inagaki. Cercozoa comprises both EF-1α-containing and EFL-containing members. **European Journal of Protistology**, 47, 24-28 (2011)
  - M. Sakaguchi, K. Takishita, T. Matsumoto, T. Hashimoto & Y. Inagaki. Tracing back the EFL evolution in the cryptomonads-haptophytes assemblage: Separate origins of EFL genes in haptophytes, photosynthetic cryptomonads, and goniomonads. **Gene**, 441, 126-131 (2009)
  - R. Kamikawa, Y. Inagaki & Y. Sako. Direct phylogenetic evidence for lateral transfer of elongation factor-like gene. **Proceedings of the National Academy of Sciences of the United States of America**, 105, 6965-6969 (2008)
- Other projects
  - R. Kamikawa, Y. Inagaki, M. Tokoro, A. J. Roger & T. Hashimoto. Split introns in the genome of Giarida intestinalis are excised by spliceosome-mediated trans-splicing. **Current Biology**, 21, 311-315 (2011)

21 other original papers and 2 review papers
In total, 38 papers in evolutionary studies for 6 years

Research collaborations (within CCS)

- Collaboration with Division of High Performance Computing System
  - Hybrid MPI/OpenMP parallelization of phylogeny programs with non-homogeneous models
  - Shota Ishikawa, PhD course student in Biological Sciences
    Co-supervised by Dr. Mitsuhisa Sato in Division of HPCS as a Master's Program student under the dual degree program
- Collaboration with Division of Computational Informatics
  - Development of a database to handle the next generation sequence data generated from diverse eukaryotic lineages
  - Yuki Nishimura, PhD course student in Biological Sciences
    Co-supervised by Drs. Hiroyuki Kitagawa and Toshiyuki Amagasa in Database Group in Division of CI
- Collaboration with Biological Function and Information Group in Division of Life Sciences
  - Prediction of tertiary structure and protein-protein interactions of translation elongation factors in eukaryotes

Research collaborations (Wet-lab collaborations outside CCS)

Wet-lab of Molecular Evolution Group:
Laboratory of Molecular Evolution of Microbes (MEM)
Faculty of Life and Environmental Sciences

-	Ken Ishida	Laboratory of Plant Systematics and Phylogeny Faculty of Life and Environmental Sciences
-	Ryoma Kamikawa	Graduate School of Global Environmental Studies Kyoto University
-	Tomoyoshi Nozaki	Laboratory of Molecular Parasitology
-	Kisaburo Nagamune	National Institute of Infectious Diseases
-	Kiyotaka Takishita	Deap-sea Ecosphere Research Team Japan Agency for Marine-Earth Science and Technology
-	Takeshi Nara	Laboratory of Molecular Parasitology Juntendo University School of Medicine

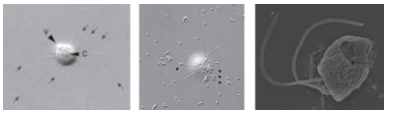
Future plans

Methodological studies in molecular phylogeny

- Simulation studies for assessing the performance of phylogenetic methods
- Development of
  - ・Phylogenetic programs for large scale ML analyses
  - ・Databases of massive NGS data from diverse eukaryotic organisms

Evolutionary studies on eukaryotic cells and their genomes

- Global eukaryotic phylogeny including:
  Microheliella maris,
  Rigifila ramosa,
  Strains SRT116, SRT127,
  SRT149, PAP020



Microheliella *Rigifila* SRT149

- Mitochondrial/Plastid evolution
- Evolution of bacterial endosymbions
- Molecular evolution of EF-1α/EFL and other molecules

Financial supports (2008-2013)

investigators	category	title	budget amount (x10³ JPY)
Hashimoto, Inagaki, Kamikawa, Nara	Grant-in-aid for Scientific Research (A) 2012-2015	Surveying novel spliceosomal components involved in *trans*-splicing in *Giardia intestinalis*: Implication for the evolution of spliceosomes	22,490+
Inagaki, Obokata, Kamikawa	Grant-in-aid for Scientific Research on Innovative Area 2011-2016	Modeling the bacterium-to-organelle transition by studying obligate endosymbiotic bacteria in diverse eukaryotic cells	69,940+
Nozaki, Hashimoto, Kuroda	Grant-in-aid for Scientific Research on Innovative Area 2011-2016	Diversity and evolution of mitochondria	106,600+
Hashimoto, Inagaki, Ishida	Grant-in-aid for Scientific Research (B) 2012-2014	Phylogenetic diversity of amitochondrial eukaryotes belonging to Fornicata	19,240
Inagaki	Grant-in-aid for Scientific Research (B) 2009-2012	Assessing a monophyletic assemblage of microbial eukaryotes including haptophytes and cryptophytes and its connection to the chromalveolata hypothesis	18,460
Inagaki	Grant-in-aid for Challenging Exploratory Research 2010-2011	Novel biflagellate TKB055 as a possible early-branch in the global eukaryotic phylogeny : studies on its morphology, transcriptome, and mitochondrial genome	3,420
Inouye, Hashimoto, Ishida, Nakayama, Inagaki, Moriya, Kikuchi	Grant-in-aid for Scientific Research (A) 2009-2012	Toward understanding the basics of environmental systems comprising microbial eukaryotes	45,110
Hashimoto	Grant-in-aid for Scientific Research (C) 2008-2010	Molecular phylogeny of Fornicata	4,940