

Hypothesis Testing with Evolutionary and Systems Biology Models

by

Seongho Ryu

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Biology

New York University

January, 2008

Bhubaneswar Mishra

© Seongho Ryu

All Right Reserved, 2008

“The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery every day. Never lose a holy curiosity.”

Albert Einstein

DEDICATION

For my mother and family, who have always support me to finish with my research, even during the most difficult times. Finally to my lovely wife, Yoonkyung Do, who helped me make it through these years. Without her help and continuous guidance, this would have never been possible.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Bud Mishra, for his role in inspiring this project, as well as his commitment to introspection, and to reflecting upon and exploring meaningful issues in Bioinformatics. I also thank my co-advisor, David Fitch. This dissertation could not have come to fruition without his help. I am also indebted to committee members, Suse Bryode, Dan Tranchina and Tim Cordozo for their ongoing guidance and support, as well as their frequent feedback at every stage of this project. I am also everlastingly grateful to my wife Yoonkyung Do, for her endless tolerance and her helping me maintain hope that I would indeed finish this project. I would also like to thank my mother, Jungja Kim, who inspired me from the beginning of my life. Finally, many thanks to Samantha and other lab members, who so patiently offered their time.

PREFACE

A new term, systems biology, is very widely used in current bioscience. Systems biology refers to a network of peripheral biological areas rather than a single well-delineated field. Bioinformatics is now an essential tool to fulfil the goal of systems biology. The model checking method, used in this study, is a computational model approach to propose specific testable hypotheses about a biological system with experimental validation. Through this method, it is possible to acquire quantitative description of cells or cell processes.

ABSTRACT

This study focuses on the analysis of dynamic computational models capable of explaining biochemical and evolutionary systems. First, we mathematically modeled the intrinsic apoptosis pathway by using a system of ordinary differential equations (an ODE model) generated by a systems biology tool, Simpathica, which is a simulation and reasoning system developed to study biological pathways. Caspase-9 is the protease that mediates the intrinsic pathway of apoptosis, a type of cell death. Activation of caspase-9 is a multi-step process that requires dATP or ATP and involves at least two proteins, cytochrome c and Apaf-1. “Model checking” based on comparing simulation data with that obtained from a recombinant system of caspase-9 activation provided several new insights into regulation of this protease. Our model predicts that the activation begins with binding of dATP to Apaf-1, which initiates the interaction between Apaf-1 and cytochrome c, thus forming a complex that oligomerizes into an active caspase-9 holoenzyme via a linear binding model with positive cooperative interaction rather than through network formation. Second, we proposed a model that explains the low conservation of rDNA intergenic spacer (IGS). The rDNA IGS subrepeats play an important role in enhancing RNA Polymerase I transcription, and yet, despite this functional role and presumed selective constraint, they show surprisingly few similarities. We observed and modeled that fast insertion-deletion rates of short mononucleotide microsatellite (or Poly(N) runs) can explain this paradox. We mathematically calculated the ideal frequencies of the Poly(N) runs

in random sequence and found their relative abundance in rDNA IGS subrepeats. Furthermore, by aligning sequences after modifying them by the drop-out method, i.e. by disregarding Poly(N) runs during the sequence aligning step, we uncovered evolutionarily shared similarities that fail to be recognized by current alignment programs. Our analysis led us to conclude that most diverse kinds of rDNA IGS subrepeats in one species must have been derived from a common ancestral subrepeat, and that it is possible to infer the evolutionary relationships among the rDNA IGS subrepeats of different species.

TABLE OF CONTENTS

DEDICATION	v
ACKNOWLEDGMENTS	vi
PREFACE	vii
ABSTRACT	viii
LIST OF FIGURES	xii
LIST OF TABLES	xiv
LIST OF APPENDICES	xv

PART I. Hypothesis Testing with Systems Biology Models

SUMMARY	2
INTRODUCTION	3
RESULTS	9
1.1 Model Derivation and Parameter Optimization	9
1.2 Of the two molecules, cytochrome c or dATP, which one initiates the formation of apoptosome?	13
1.3 Does cytochrome c affect caspase-9 activation after the apoptosome is formed?	17
1.4 Is there cooperative interaction during the formation of Apaf-1 complex?	19
1.5 Is binding of procaspase-9 to the apoptosome cooperative?	22
1.6 Is free caspase-9 active?	27

MATERIALS AND METHODS	32
PART II. Hypothesis Testing with Evolutionary Models	
SUMMARY	49
INTRODUCTION	51
MATERIALS AND METHODS	54
RESULTS	59
2.1 Patterns of Poly(N) in subrepeats of the IGS region	59
2.2 Drop-out algorithm	70
2.3 The drop-out method and its efficacy in revealing similarity among different types of subrepeats from a species	86
2.4 Usage of drop-out alignment method to compare subrepeats between different species	95
2.5 Usage of drop-out alignment method to find novel conserved sequences between species	102
DISCUSSION	106
APPENDICES	112
BIBLIOGRAPHY	116

LIST OF FIGURES

Fig. 1.1 Caspase-3 activation as described by a conventional diagram and in Simpathica	10
Fig. 1.2 Two models for initiation of caspase-9 activation and measurement of Caspase-3 activities	14
Fig. 1.3 Test a second role of cytochrome c in intrinsic apoptotic pathway besides the formation of Apaf-1 oligomer	18
Fig. 1.4 Cooperative binding of Apaf-1 can explain the rate of caspase-3 activation	21
Fig. 1.5. Comparison of cooperative interactions among procaspase-9 during the formation of holoenzyme complex	25
Fig. 1.6. Holoenzyme-bound caspase-9 is enough to activate caspase-329	
Fig. 1.7 Revised caspase-9 dependent intrinsic apoptotic pathway	31
Fig. 2.1 Biased base composition of the rDNA IGS and subrepeats	65
Fig. 2.2 The comparison between the Poly(N) base composition of subrepeats changes according to its frequency	69
Fig. 2.3 Sequence alignment among four subrepeats of the <i>Xenopus</i> IGS region.	88
Fig. 2.4 Sequence alignment between two subrepeats of <i>D. melanogaster</i> IGS region	92
Fig. 2.5 Sequence alignment between the 63bp and the 141bp subrepeats of tomato	94

Fig. 2.6 Multiple alignments among different subrepeats from various species after dropping out Poly(N)s	98
Fig. 2.7 Multiple alignments among many subrepeats from various species	101
Fig. 2.8. Dot-plot comparisons of whole rDNA IGS sequences from two <i>Drosophila</i> species	104

LIST OF TABLES

Table 1.1 The element reactions and kinetic parameters used in the simulation.	12
Table 1.2 Variable definitions	37
Table 1.3. Differential equations, variable definitions and default parameters	38
Table 2.1. List of species and rDNA IGS subrepeats used in this study	55
Table 2-2. The dynamic programming recurrence equation	74
Table 2.3. Comparing similarity values among long rDNA IGS subrepeats from the plant species before and after dropping out Poly(N)	96

LIST OF APPENDICES

APPENDIX A:	One-Piece Drop-out Algorithm	112
-------------	------------------------------	-----

PART I. Hypothesis Testing with Systems Biology Models

(Mathematical Modeling of the Intrinsic Pathway of Apoptosis using
Simpathica, a Systems Biology Tool)

SUMMARY

Caspase-9 is the protease that mediates the intrinsic pathway of apoptosis, a type of cell death. Activation of caspase-9 is a multi-step process that requires dATP or ATP and involves at least two proteins, cytochrome c and Apaf-1. In this study, we mathematically model caspase-9 activation by using a system of ordinary differential equations (an ODE model) generated by a systems biology tool Simpathica - a simulation and reasoning system, developed to study biological pathways. “Model checking” based on comparing simulation data with that obtained from a recombinant system of caspase-9 activation, provided several new insights into regulation of this protease. The model predicts that the activation begins with binding of dATP to Apaf-1, which initiates the interaction between Apaf-1 and cytochrome c, thus forming a complex that oligomerizes into an active caspase-9 holoenzyme via a linear binding model with cooperative interaction rather than through network formation.

INTRODUCTION

Systems biology aims to understand the fundamental principles governing biological systems within a mathematical framework. There are many advantages to using a mathematical framework: the most important being that it would then be relatively easy to translate biological systems into *in silico* models that can be manipulated computationally and symbolically in a manner that is open simply impossible with any *in vitro* or *in vivo* models. Furthermore, it opens up the opportunities to adapt many algorithmic and analysis techniques originally devised to study natural or engineered dynamical systems that are often represented as discrete, continuous or hybrid models and analyzed using classical or modal logical frameworks.

In engineering sciences, such a consilience of model building and model checking approaches has resulted in robust and well validated engineered systems in spite of the fact that their design could often evolve uncontrollably in complexity to become too incomprehensible too quickly (e.g., internet or power-grid), even to the team of engineers who may have originally designed the system. But the biological systems bring up even further obstacles: often we do not know all the components participating in the process, nor possess the absolute knowledge of all the underlying interactions, complete with their kinetic parameters. To overcome these obstacles, it is important to develop system tools capable of establishing models with symbolic

methods, analyzing them by mathematical approaches, evaluating the models with numerical simulations, and suggesting the best model system to recapitulates the biological system of interest. Recently, our group developed an automated software platform, dubbed Simpathica Model Checker, which provides “in one package” a set of the necessary systems biology tools to carry out studies of enzymatic metabolic pathways (Mishra et al., 2005). In this study we applied this platform to model the set of biochemical processes that are thought to be occurring during apoptosis.

Apoptosis is a type of cell death that removes malfunctioning or damaged cells (Evan and Littlewood, 1998; Riedl and Shi, 2004). Apoptosis is triggered when DNA is damaged, cells are detached from neighbors, growth factor is deprived, or cells are infected. Accordingly, a failure of apoptosis can contribute to diseases, such as cancer, while uncontrolled activation of the apoptotic machinery can cause unwanted loss of cells such as that occurring in neurodegenerative diseases. Apoptosis is executed by two main pathways, intrinsic, which responds primarily to intracellular stress, such as caused by chemotherapy, or extrinsic, which is activated by agonists of so-called death receptors (Cain, 2003). The proteins involved in these pathways have been characterized, but how these proteins interact to ensure that apoptosis is prevented until needed but then executed quickly remains unclear. It is likely, however, that this regulation involves quantitative changes in concentrations and activities of these proteins.

Both pathways of apoptosis can be described as a network that regulates activation of caspases, the proteases that disassemble the cell. Caspases (cysteine aspartate-specific proteases) is a family of proteins that form the execution part of the apoptotic machinery. The extrinsic pathway activates caspase-8, while intrinsic pathway activates caspase-9; either of these proteases can activate caspase-3, which does most of cell disassembly by cleaving a set of proteins. Caspase-9 functions as a holoenzyme, in which this protease is a catalytic subunit that is regulated by an oligomer of Apaf-1, also known as the apoptosome (Jiang and Wang, 2004).

Apaf-1 is a 130-kDa constitutively expressed protein that includes a caspase recruitment domain (CARD), a nucleotide-binding domain and 13 WD-40 repeats (Zou et al., 1997; Zou et al., 1999). Apaf-1 is oligomerized into the apoptosome following binding to cytochrome c in a process that requires hydrolysis of ATP or dATP (Kim et al., 2005; Yu et al., 2005). Because of experimental difficulties with studying the apoptosome activation, it is uncertain what sequence of reactions leads exactly to caspase-9 activation, or even how this protease is activated.

Cytochrome c is very important molecule in oxygen-rich earth environment, but at the same time, it is key molecule to initiate cell death. Cytochrome c is very conserved molecules from yeast cells to human cells because of its essential function(s). The well-known function of cytochrome c is to shuttle electrons through

the last step of aerobic energy production. However, cytochrome c has a negative function as well. Cytochrome c and the mitochondria play a central role in apoptosis, signaling for programmed cell death.

We hope that modeling processes leading to caspase-9 activation may help to learn how to regulate caspase-9 activation for therapeutic needs. In particular, the modeling approach could predict the outcome in a perturbed system where the balance between cell survival and cell death is compromised. Modeling can also facilitate determination of unknown factors involving the apoptotic pathway, and be a platform for exchanging knowledge and storing information. Finally, a model of caspase-9 activation can be eventually merged with other models modularly and hierarchically to form a functional model of the cell.

Since Varner and colleagues proposed the first mathematical model of caspase activation (Fussenegger et al., 2000), several additional details have been proposed to enhance mathematical models of apoptosis by combining other major elements thought to be involved in apoptosis (Legewie et al., 2006; Nakabayashi and Sasaki, 2006; Stucki and Simon, 2005). For instance, there had existed earlier models to provide frameworks analogous to ours, and account for the various interactions that can affect apoptosis; however, the present study attempts to go much further in filling in many details of the intrinsic pathway that needed a clearer and more detailed

description. As an example, Nakabayashi and Sasaki had previously mathematically modeled apoptosome assembly with the network interaction (Nakabayashi and Sasaki, 2006); however, in contrast to a pure simulation study as theirs, we have been able to combine model building with model checking using novel recombinant experimental data and automated systems biology tools.

This study not only proposes a mathematical model of caspase-9 activation, but also describes a platform that can be used even by a novice user to apply, test, extend, or modify such models. This platform is based on NYU's Simpathica Model Checker (SMC) that is part of the VALIS software environment (Paxia et al., 2002). SMC was designed for modeling, simulation, and reasoning, and is capable of effectively manipulating large, complex, and highly detailed biochemical systems (Mishra et al., 2005). Importantly, Simpathica can generate all the necessary differential equations starting from the user-defined textual or graphical descriptions, which allows even a user who is unfamiliar with this mathematical approach to apply it effortlessly. Simpathica allows users to modify an existing model, modularly integrate with established or hypothetical models, or search over a family of plausible models, through a simple and efficient Graphical user interface (GUI) using multi-scripting facilities of VALIS. In summary, Simpathica allows users to construct and simulate models of metabolic, regulatory, and signaling networks and then to analyze their behavior with equal ease.

Here, we used Simpathica to model the intrinsic pathway and checked the model by comparing modeling predictions with experimental data.

RESULTS and DISCUSSION

1.1 Model Derivation and Parameter Optimization

Our initial modeling is based on the current hypothesis of intrinsic apoptosis, which is initiated by the release of cytochrome c from mitochondria (Jiang and Wang, 2004). In this model (Fig.1.1A), the released cytochrome c binds Apaf-1 monomer and the binding makes Apaf-1 active (step 1 in table 1 and fig. 1.1A, eq. 13). The active Apaf-1 then binds to dATP and forms an oligomer, usually referred to as apoptosome (step2, eq. 14). The oligomerization is simplified in a Michaelis-Menten reaction as reported ((Nakabayashi and Sasaki, 2006), step 3) or in a series of Michaelis-Menten reactions (eq. 17-22). The resulting apoptosome recruits pro-caspase-9 to form the caspase-9 holoenzyme (eq. 23). The holoenzyme becomes active (eq. 24, step 4) and can catalyze the activation of caspase-3 (step 5), whose activity is monitored by the increasing concentration of processed peptide substrate DEVD-Afc (step 6). The active holoenzyme can also dissociate into the apoptosome and free processed caspase-9 (eq. 25, step 7).

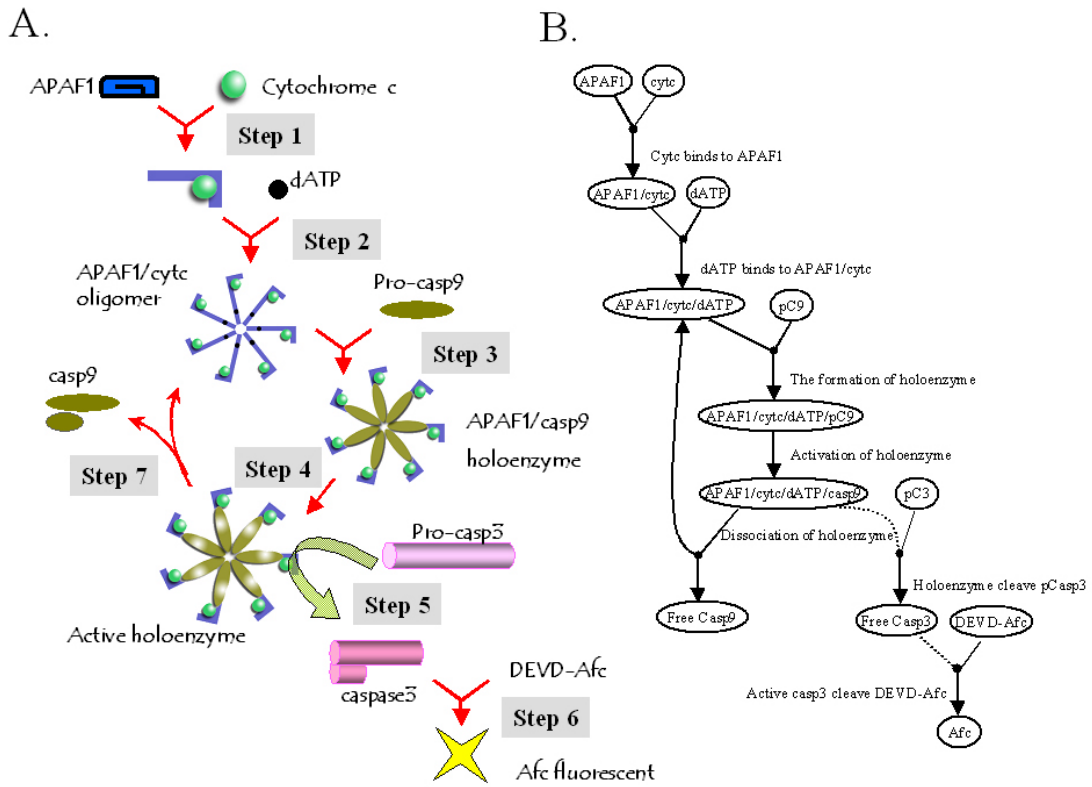


Figure 1.1. Caspase-3 activation as described by a conventional diagram (A) and in Simpathica (B). Dashed line indicates enzymatic reactions.

Since any modeling requires quantitative parameters of modeled processes, we assembled published kinetic parameters of processes involved in caspase-9 activation, determined these parameters experimentally using an *in vitro* or a recombinant system, or postulated them (Table 1.1). We compared the results of simulation with the published experimental data (Rodriguez and Lazebnik, 1999), and changed the kinetic constants used in the simulation to reconcile the differences between the simulated and experimentally observed processes.

Remarkably, changing most kinetic parameters in the simulation, even by several orders of magnitude, had little effect on caspase-9 activation. However, changing some parameters within a relatively narrow range significantly changed the kinetics of active caspase. Changing the K_m and k_2 of the active holoenzyme (step 5) had the largest effect on caspase-9 activation, suggesting that these parameters are the key determinants of caspase-9 regulation. For example, decreasing the K_m of the active holoenzyme increases the rate of pro-caspase-9 processing. On the other hand, decreasing the dissociation constant of the active holoenzyme (step 7) accelerates the cleavage and activation of both caspase-3 and caspase-9.

Using the optimized simulation, we posed several specific questions regarding intrinsic apoptosis pathway.

Table 1.1. The element reactions and kinetic parameters used in the simulation.

Step	Species or reactions (eq. number)	Values observed in experiments	Values used in simulation
0	Nucleotide (D)	[ATP]=2~10x10 ⁻³ M (Zubay, 1993) ^C	5 x10 ⁻³ M
	Cytochrome c (C)	(i)1 [cytc]=1.1x10 ⁻⁶ M ^E	1.2 x10 ⁻⁶ M
	Apaf-1 (A)	[APAF1]=1.5~3x10 ⁻⁹ M (Fearhead et al., 1998; Zou et al., 1997) ^E	4 x10 ⁻⁹ M
	Caspase-9 (C9)	[casp9]=2 x10 ⁻⁸ M (Stennicke et al., 1999) ^E	2 x10 ⁻⁸ M
	Caspase-3 (C3)	[casp3]=1 x10 ⁻⁷ M (Stennicke et al., 1998) ^E	1.5 x10 ⁻⁸ M
	Caspase-3 substrate	[DEVD-Afc]=4 x10 ⁻⁵ M (Rodriguez and Lazebnik, 1999) ^E	4 x10 ⁻⁵ M
1	C + A → A/C (eq.13)	$k_{on}=10^7 \text{ M}^{-1}\text{s}^{-1}$, $k_{off}=10^{-4} \text{ s}^{-1}$ (Purring et al., 1999) ^R $K_A=4 \times 10^7 \text{ M}^{-1}$ (Purring-Koch and McLendon, 2000) ^R	$k_1=6 \times 10^{17} \text{ M}^{-1}\text{s}^{-1}$, $k_{-1}=6 \times 10^3 \text{ s}^{-1}$ $k_2=2 \text{ s}^{-1}$
2	A/C + D → A/C/D (eq.14)	$K_D=1.72 \times 10^{-6} \text{ M}$ (Jiang and Wang, 2000) ^R	$k_1=1 \times 10^{12} \text{ M}^{-1}\text{s}^{-1}$, $k_{-1}=1.72 \times 10^3 \text{ s}^{-1}$ $k_2=5 \times 10^{-1} \text{ s}^{-1}$
3	A/C/D + P9 → A/C/D/P9 (eq.17~22)	(ii)	$k_1=2 \times 10^{11} \text{ M}^{-1}\text{s}^{-1}$, $k_{-1}=2 \times 10^4 \text{ s}^{-1}$ $k_2=1 \times 10^2 \text{ s}^{-1}$
4	A/C/D/P9 → A/C/D/C9 (eq.24)	(iii)	$k_2=3 \text{ s}^{-1}$
5	A/C/D/C9 P3 → C3 (eq.27)	$K_i=1.08 \times 10^{-7} \text{ M}$ (Legewie et al., 2006) ^R (iv)	$k_1=6 \times 10^9 \text{ M}^{-1}\text{s}^{-1}$, $k_{-1}=1.08 \times 10^3 \text{ s}^{-1}$, $k_2=2 \times 10^1 \text{ s}^{-1}$
6	C3 V → F (eq.28)	$K_M=9 \times 10^{-6} \text{ M}$ $k_2=7.5 \times 10^{-1} \text{ s}^{-1}$ (Moretti et al., 2002) ^R	$k_1=5 \times 10^{12} \text{ M}^{-1}\text{s}^{-1}$, $k_{-1}=5 \times 10^4 \text{ s}^{-1}$ $k_2=9 \times 10^1 \text{ s}^{-1}$
7	A/C/D/C9 → A/C/D + C9 (eq.25)	(v)	$k=0.5 \text{ s}^{-1}$

*All abbreviations are listed in table 2.

C: cell, E: extract system, R: recombinant system.

(i) Determined in the 293 extract by cytochrome c ELISA kit (R&D system).

(ii) Postulated, half of caspase-9 can be recruited in holoenzyme in 5 minutes with the $K_i > 10000 \text{ nM}$ (Inhibition of Caspase-9 by caspase-9 C285A) (Ryan et al., 2002)

(iii) Postulated. 20nM of caspase-9 can be completely cleaved in 20 minutes.

(iv) Inhibition of Caspase-9 by peptide aldehydes (Garcia-Calvo et al., 1998)

(v) Postulated.

1.2 Of the two molecules, cytochrome c or dATP, which one initiates the formation of apoptosome?

The current model of Apaf-1 activation assumes that Apaf-1 is bound to dATP, which is hydrolyzed upon binding of Apaf-1 to cytochrome c. The resulting dADP remains bound to the Apaf-1-cytochrome c complex and then is exchanged by an undefined mechanism for a molecule of dATP, thus producing a complex that is oligomerized into the apoptosome. Therefore, in this model the binding of cytochrome c to Apaf-1 is the triggering event for the formation of apoptosome and the consequent caspase-9 activation (Kim et al., 2005; Yu et al., 2005).

However, previous studies provided evidence that cytochrome c binds Apaf-1 independently of the presence of free dATP (Zou et al., 1997), that binding to cytochrome c induces binding of Apaf-1 to dATP (Cain et al., 2000; Jiang and Wang, 2000), and that hydrolysis of dATP by Apaf-1 is continuous and precedes binding of cytochrome c (Zou et al., 1999). The possible discrepancies between these observations and the model suggest that the exact sequence of reactions that result in the active apoptosome is yet to be established. We used our simulation to compare the effect on caspase-9 and caspase-3 activation of two possible initiating steps: binding of cytochrome c induces binding of dATP (Fig. 1.2A), or binding of dATP induces binding of cytochrome c (Fig. 1.2B).

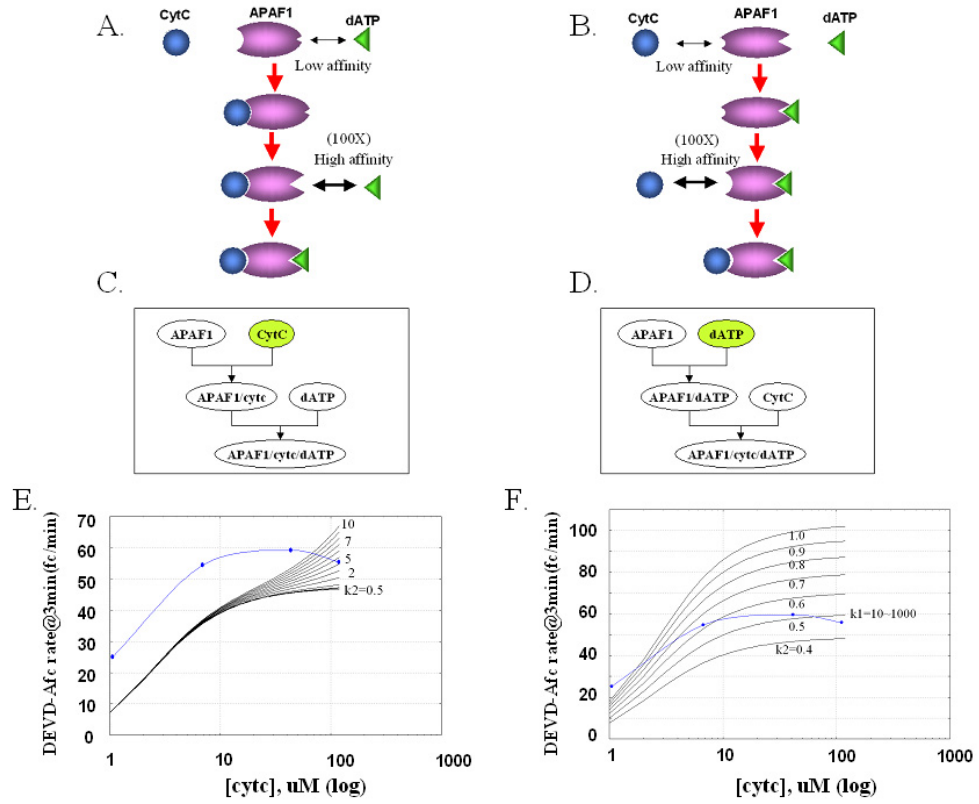


Figure 1.2. Two models for initiation of caspase-9 activation and measurement of Caspase-3 activities using recombinant system and comparison with simulation data from two apoptosis initiation models. (A) and (B) are the schematic diagrams indicating the first binding molecule to the Apaf-1 and (C) and (D) are the diagrams generated by the simulation program, Simpathica. We use notation 100X to indicate that rate constant of the forward reaction is increased by a scale factor of 100 (unit is $M^{-1}s^{-1}$). (E) and (F) are the corresponding graphs from the simulation. Caspase-3 activities were calculated at various concentration of cytochrome c up to 120uM after two minutes. Both graphs were compared with the recombinant experiment (blue lines). Caspase-3 activities were measured at 7 different concentration of cytochrome c, from 0 to 120 μ M after three minutes.

Simulation of the first model indicated an unstable fluctuation of caspase-3 activity. To test the robustness of the models, we also conducted extensive perturbation analysis by simulating the model with perturbed rate constants for the binding reaction of cytochrome c with Apaf-1. In this model, decreased k_1 causes no change of caspase-3 activity in lower concentration of cytochrome c (Fig. 1.2E). This result indicates that the formation of holoenzyme in this model is not sensitive to the concentration of cytochrome c. Thus, small change of concentration may not control the intrinsic apoptotic system. However, the simulation of the second model, ATP model, produced a stable increase of caspase-3 activity at the broad range of cytochrome c concentration regardless of changing rate constants. This increase recapitulated the shape assumed by the experimentally observed data in the recombinant system (Fig. 1.2F). Therefore, the simulation favored the model in which the first step of apoptosome formation is binding of dATP to Apaf-1 –a process that then facilitates subsequent binding of cytochrome c.

1.3 Does cytochrome c affect caspase-9 activation after the apoptosome is formed?

Cytochrome c triggers oligomerization of Apaf-1, but is still detected in the holoenzyme, which raises the question as to whether this protein is redundant in the apoptosome or has some specific but unidentified functions. Therefore, we compared two simulations. In one, cytochrome c is involved only in the formation of Apaf-1 complex, while in the other cytochrome c is further assumed to participate in holoenzyme activation, accompanied by increasing reaction rate. We found no detectable difference between the two models except in the time at which caspase-3 activity reaches its maximum value (Fig. 1.3). Thus, we concluded that cytochrome c is unlikely to be involved in the caspase-9 activation. However, this analysis does not exclude the possibility that cytochrome c in the complex has a role in stabilizing the holoenzyme complex.

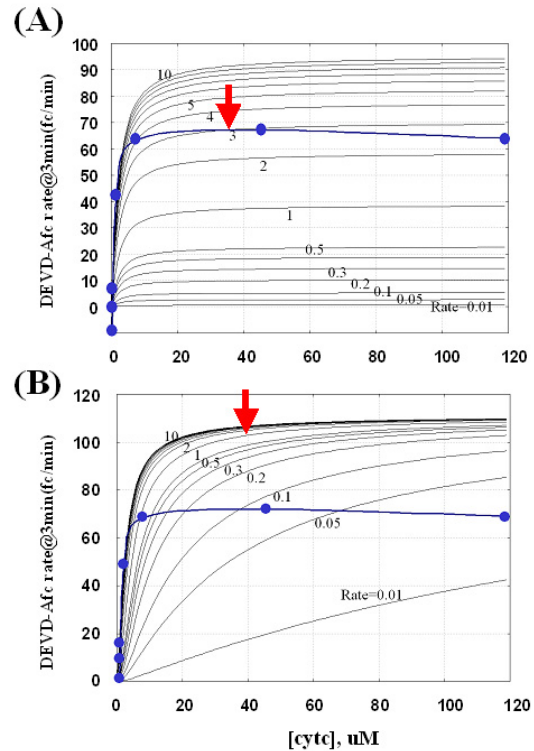
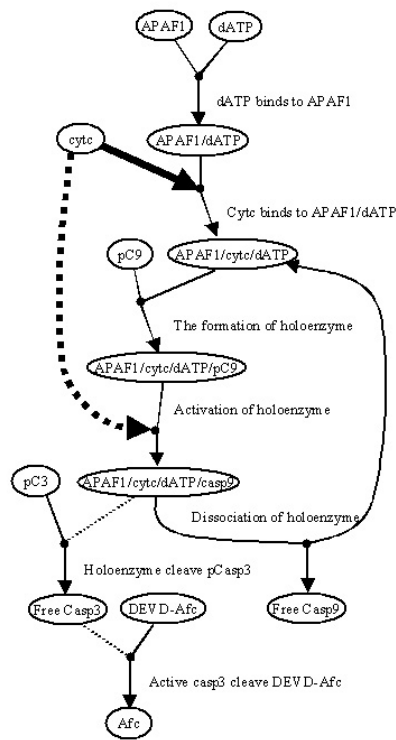


Figure 1.3. Test a second role of cytochrome c in intrinsic apoptotic pathway besides the formation of Apaf-1 oligomer. Left panel shows the modeling pathway view in Simpathica. (A) arrow indicates the time point at which caspase-3 activity reaches the maximum activity in a graph corresponding to the model in which cytochrome c is only involved in the formation of Apaf-1 complex. (B) arrow indicates the time point at which caspase-3 activity reaches the maximum activity in a graph corresponding to the model in which cytochrome c has a dual role, controlling how Apaf-1 complex (arrow with solid line in the diagram) and enzymatic activity facilitates holoenzyme activation (arrow with broken line in the diagram). Red arrow indicates the time point.

1.4 Is there cooperative interaction during the formation of Apaf-1 complex?

A key step in activation of caspase-9 is oligomerization of Apaf-1. How this oligomerization occurs is not known. The rapid rate of the oligomerization and the number of the subunits in the oligomer suggests a possibility of cooperative binding among the subunits. Recently, Nakabayashi and Sasaki mathematically modeled apoptosome assembly with a quadratic network interaction framework that did not involve cooperative binding (Nakabayashi and Sasaki, 2006).

To test whether cooperative binding can also explain how Apaf-1 is oligomerized, we simulated the effect of cooperative binding on the activation of caspases-9 and -3, we assumed that Apaf-1 oligomerization proceeds stepwise (equation (17) through (22)). We created two simulation models with a linear network interaction framework. In one of the models, each Apaf-1 complex subunit has the same binding rate as previously bound subunits ($\gamma_1 = \gamma_2 = \gamma_3 \dots = \gamma_n$; γ_n is binding rate of n th Apaf-1 complex), while in the other model the successive Apaf-1 complex subunits have monotonically increasing binding rates ($\gamma_1 \ll \gamma_2 \ll \gamma_3 \dots \gamma_n$).

For comparison, we also simulated the Nakabayashi and Sasaki's complex network model (Nakabayashi and Sasaki, 2006) in which each Apaf-1 monomer binding to the oligomer has the same binding rate, but the monomers and oligomers

participate in complex network interactions. For example, an Apaf-1 trimer can bind to a tetramer to form the heptameric apoptosome, or bind to a dimer, to form an intermediate pentamer, and so on. In this model, a total of 21 individual interactions would be necessary to connect all components needed in the formation of a heptamer (Fig. 1.4C).

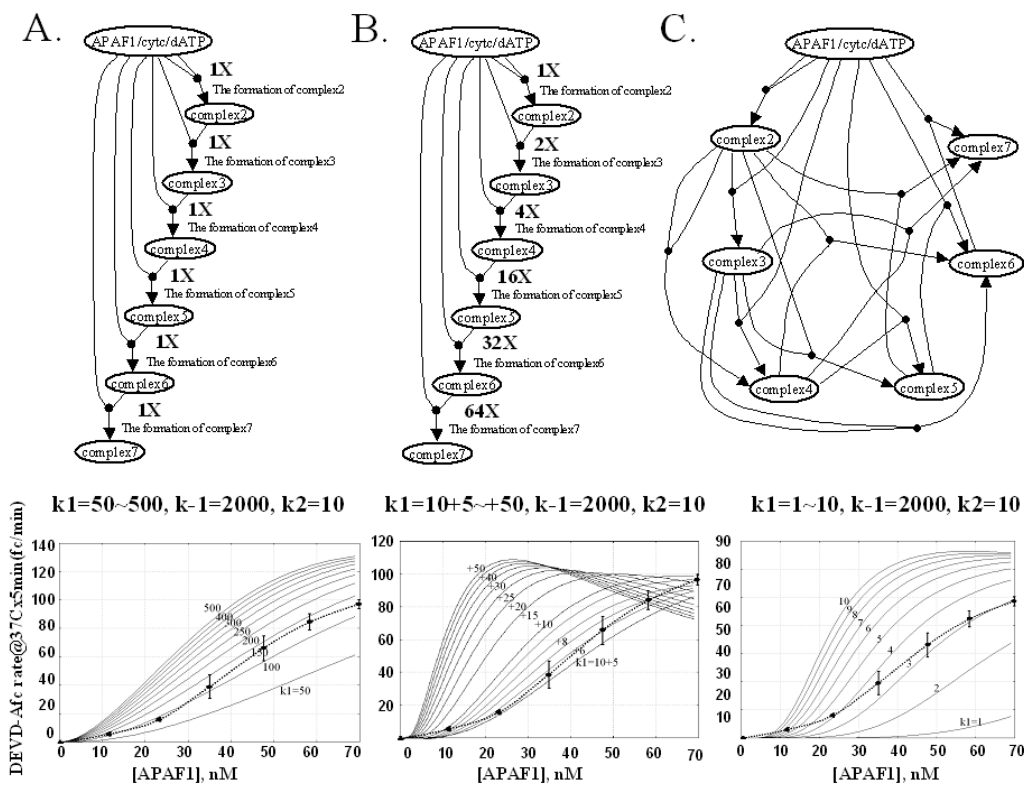


Figure 1.4. Cooperative binding of Apaf-1 can explain the rate of caspase-3 activation.

Top panel shows the modeling pathways view in Simpathica and bottom panel shows the results from simulation and *in vitro* experiment. (A) a linear network model with non-cooperative interactions; (B) a linear network model with cooperative interactions; and (C) a complex network model with non-cooperative interactions. Blue lines indicate graphs from the recombinant experiment. Caspase-3 activities were measured at 7 different concentration of Apaf-1, from 0 to 70nM. Caspase-3 activities were calculated at various concentration of Apaf-1 up to 70nM (dashed lines).

In the linear model the increase of caspase-3 activity was exponential if no cooperativity was assumed (Figure 4A), and sigmoidal if cooperativity was introduced (Figure 4B). The Nakabayashi and Sasaki model produced a sigmoidal curve that was initially nearly linear (Figure 4C). To determine which of the curves fits better the experimental results, we calculated the mean square error (MSE, see Theoretical basis for Simulation) between experimental and simulation data using a least square distance function, and concluded that the model based on a linear network with cooperative interactions (Fig. 4B) fit the best. We also tested the robustness of the models by varying the binding rate constants. As we expected, increased binding rates elevate the holoenzyme activity. However, their increasing pattern was not changed and overall shape remained unchanged. These simulation results indicate that modified parameters do not change the patterns of the dynamics. Therefore, we may reasonably conclude that there exist positive cooperative interactions during the Apaf-1 complex formation.

1.5 Is binding of procaspase-9 to the apoptosome cooperative?

Caspase-9 functions as a holoenzyme in which this protease and Apaf-1 are present in 1:1 ratio (Acehan et al., 2002). Thus, total seven caspase-9 molecules can exist in holoenzyme complex even though it is not clear how this oligomerization occurs. The simulation suggesting a cooperative binding among Apaf-1 complex subunits (Figure 4) led us to test whether binding of procaspase-9 to the oligomer is

also cooperative. We compared two alternative models. In one model, each procaspase-9 molecule binds with the same binding constant as the previous procaspase-9 during the formation of the holoenzyme ($\delta_1 = \delta_2 = \delta_3 \dots = \delta_7$; δ_n is binding rate of n th procaspase-9 molecule), while in the other alternative model the rate of binding of each successive procaspase-9 molecules increases monotonically ($\delta_1 \ll \delta_2 \ll \delta_3 \dots \ll \delta_7$).

We found that assuming cooperative binding of caspase-9 to the Apaf-1 oligomer could not by itself explain the sigmoidal shape in the rate of caspase-3 activation observed experimentally (Fig.5 left and right panel). Moreover, positive cooperative binding of caspase-9 to the Apaf-1 oligomer could accelerate the holoenzyme activity in the higher concentration of Apaf-1 relative to the activity measured by the experimental data. We also tested the robustness of the models by perturbing binding rate constants. However, since the overall shape remained the same, we concluded that caspase-3 activation involves cooperative interaction among Apaf-1 molecules during the formation of Apaf-1 multimeric complex in low concentration of Apaf-1 (Fig.5 bottom panel).

Based on the previous results, we proposed that seven Apaf-1 complex subunits interact cooperatively with each other while forming the apoptosome, and that this positive cooperation can explain the rapidity of caspase-9 activation during

apoptosis, especially considering that assuming positive cooperativity in the binding of caspase-9 to the apoptosome failed to affect the rates of caspase-3 activation

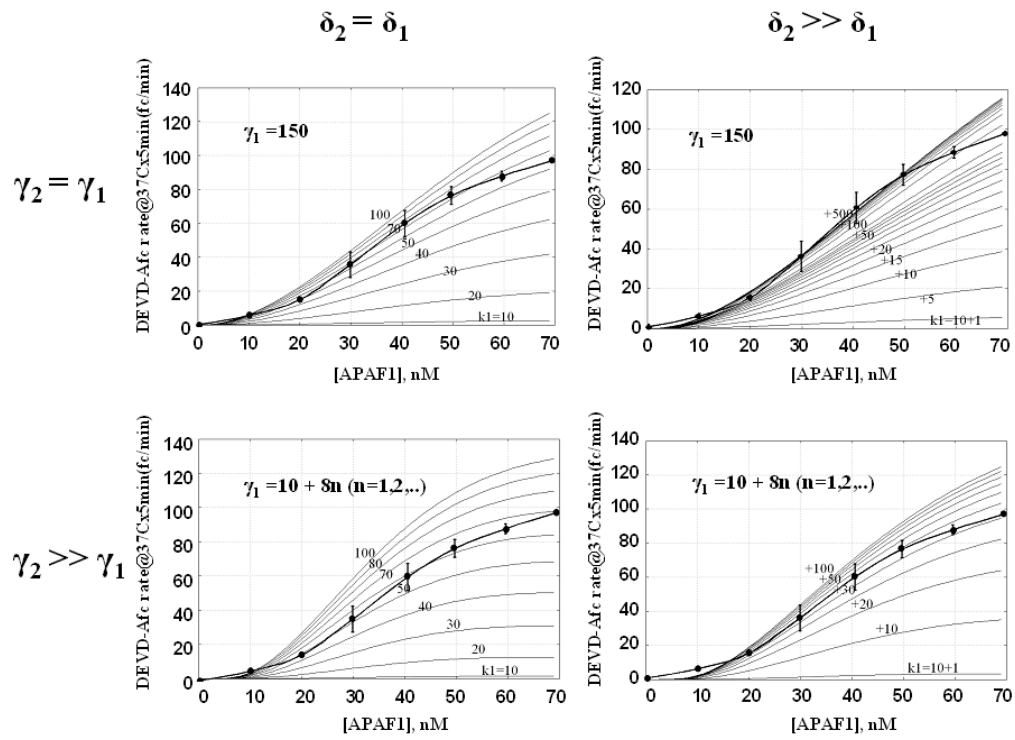
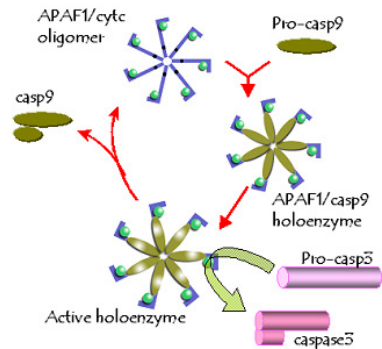


Figure 1.5. Comparison of cooperative interactions among procaspase-9 during the formation of holoenzyme complex and cooperative interaction among Apaf-1 complex subunits during the formation of Apaf-1 multimeric complex. Left panel exhibits an absence of cooperative interaction among procaspase-9 during the formation of holoenzyme complex. Right panel exhibits a cooperative interaction among procaspase-9 during the formation of holoenzyme complex. Top panel exhibits no cooperative interaction among Apaf-1 complex subunits during the formation of Apaf-1 multimeric complex. Bottom panel exhibits a cooperative interaction among Apaf-1 complex subunits during the formation of Apaf-1 multimeric complex (unit of γ and δ is $M^{-1}s^{-1}$). In all simulation, caspase-3 activities were calculated at various concentration of Apaf-1 up to 70nM. Bold lines indicate graphs from the recombinant experiment. Caspase-3 activities were measured at 7 different concentration of Apaf-1, from 0 to 70nM.

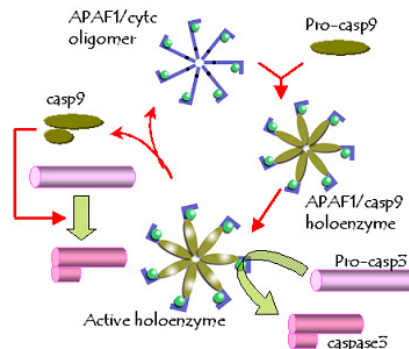
1.6 Is free caspase-9 active?

The experimental evidence indicates that caspase-9 is fully active only when it is bound to Apaf-1 (Rodriguez and Lazebnik, 1999; Stennicke et al., 1999), while earlier studies suggested that processed caspase-9 can also be active (Cain et al., 2000; Zou et al., 1999). We simulated these two possibilities (Fig. 1.6) and found that the first model fits the experimental findings better. Simulation study shows that if free caspase-9 also has any activity, holoenzyme activity in high concentration of Apaf-1 should have been much higher than what we observed in experiments. When we compared the activity of free caspase-9 against that of complex-bound caspase-9, we observed that holoenzyme activity was increased by almost 50%, in higher concentration of Apaf-1. However, if we examine the activities at a low level of free or complex-bounded caspase-9s, we observed both holoenzyme activities to be of similar magnitudes. Therefore, we concluded that free caspase-9 has none or very little influence. Our simulation data confirmed the result from recent experimental data that caspase-9 is fully active only when it is bound to Apaf-1 (Rodriguez and Lazebnik, 1999), and that, in contradiction to earlier mathematical models, the apoptotic process is maintained primarily by caspase-3 or caspase-9 (Fussenegger et al., 2000; Legewie et al., 2006; Stucki and Simon, 2005).

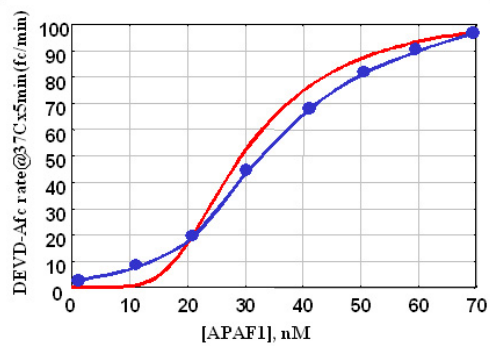
A.



B.



C.



D.

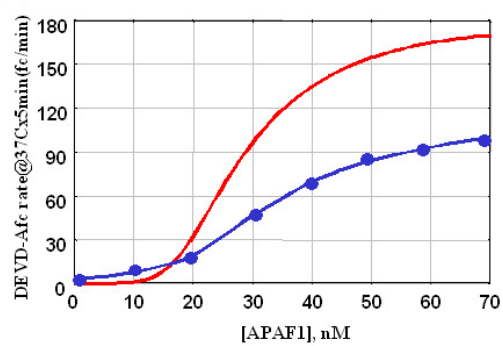


Figure 1.6. Holoenzyme-bound caspase-9 is enough to activate caspase-3. (A) shows the schematic diagrams which implicitly assume a role only for holoenzyme-bound caspase-9 in activating caspase-3. (B) shows the schematic diagrams which assume that both holoenzyme bound caspase-9 and free caspase-9 jointly activate caspase-3. (C) shows the output of a simulation model from panel A. (D) shows the output of a simulation model from panel B. In all simulation, caspase-3 activities were calculated at various concentration of Apaf-1 up to 70nM. Dashed lines indicate graphs from the recombinant experiment.

Based on simulation data, we reconstructed the model for the intrinsic apoptotic pathway (Fig.1.7). Based on the analysis from this study, we concluded that the best target molecule to control apoptosis is not necessarily initiator caspase such as caspase-9 but a holoenzyme complex. Although the behavior of caspase-3 activity is mainly caused by cooperative binding of Apaf-1 subunits, cooperative binding of procaspase-9 still can affect caspase-3 activity in high concentration of Apaf-1. It is interesting that the simulation data from this study based on cooperative interactions almost matched with *in vitro* experimental data for all data points. This result supports the main claim of this study that all components and their interactions used in this simulation study are structured appropriately in their ability to use the experimental data to distinguish among a family of plausible models.

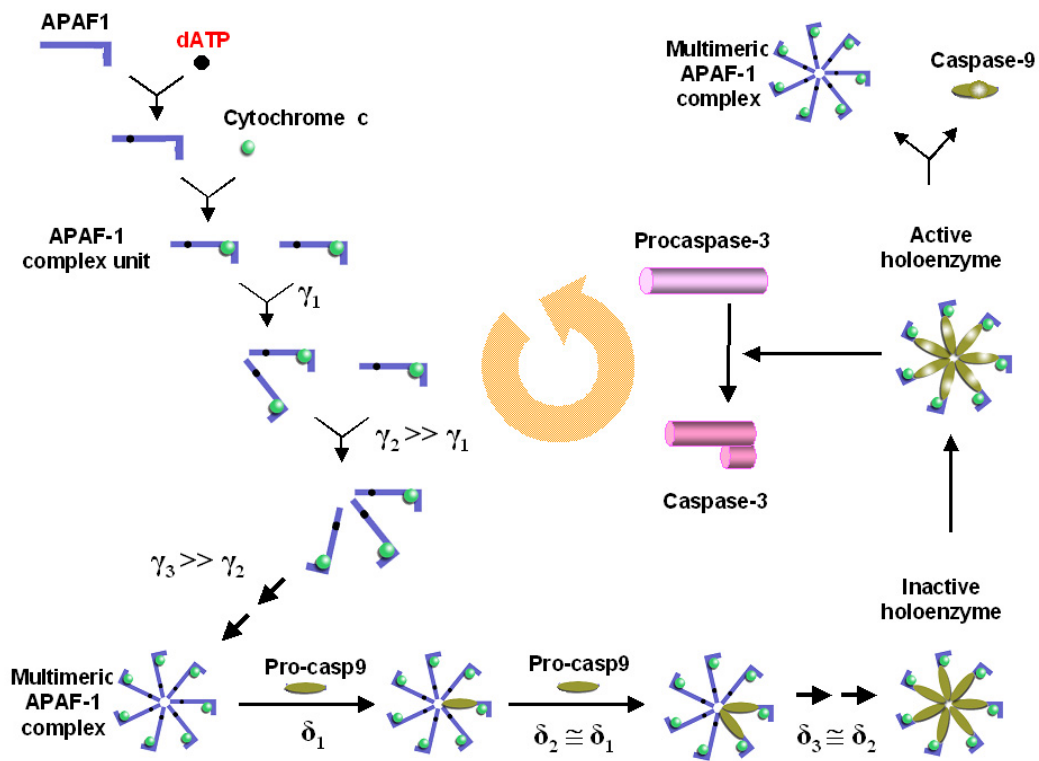


Fig.1.7 Revised caspase-9 dependent intrinsic apoptotic pathway. Variable γ denotes rate constant for the binding between Apaf-1 complex units during the formation of apoptosome and δ denotes rate constant for the binding between multimeric Apaf-1 complex and procaspase-9.

Materials and Methods

Simulation program, Simpathica

The model equations and a graphical user interface (GUI) were generated, using Simpathica Model Checker (developed by NYU Bioinformatics Group, New York University, NY). This program is publicly available at NYU Bioinformatics Group website: <http://bioinformatics.nyu.edu/Software/index.shtml>.

Theoretical basis for Simulation

Simpathica is an advanced integrated systems biology tool for simulating and reasoning about biological processes (Mishra et al., 2005). Simpathica is able to model large, modular and hierarchical biochemical pathways without making overly simplifying assumptions. For instance, at the lowest level, Simpathica may model a typical metabolic reaction modulated by an enzyme using the classical Michaelis-Menten's formulation. The parameters of such an equation would be the constants K_m (Michaelis-Menten Constant) and V_{max} (maximum velocity of a reaction). Simpathica follows a modular scheme, much as in S-system ((Voit, 1991; Voit, 2000), to provide in a convenient manner a description of a biochemical pathway as a composition of several primitive reaction modules, which can be automatically translated into a set of

ODE's with additional algebraic constraints. Simpathica and XS-system (an extension of the basic S-System, (Antoniotti et al., 2003a; Antoniotti et al., 2003b; Antoniotti et al., 2002)) retains this modular structure while allowing for a far richer set of modules and constraints. The Simpathica architecture is composed of two main modules and several ancillary ones. The first main module is a graphical front end that is used to construct and simulate the networks of ODE's that are part of the model being analyzed. Simpathica uses, among others, the SBML format (System Biology Markup Language, www.sbml.org) for exchanging model descriptions. The second module, XSSYS is an analysis module based on a branching time temporal logic, that can be used in formulating questions about the behavior of a system, represented as a set of traces (time course data) obtained from wet-lab experiments or computer simulations.

Comparison of simulation model predictions against experimental data

In this study, our model checking approach consisted of numerical comparison between simulation (*in silico*) data, $\mathfrak{S}_s(U, T)$, and experimental (*in vitro*) data, $\mathfrak{S}'_v(U, T)$, leading to a better characterization of correct apoptosis model. Based on this initial ODE model, a suitable Kripke structure can be created to check more complex modal logic queries within Simpathica (data not shown).

In general, we may assume that we use an *in silico* model, expressed by a

system of ODE's

$$\frac{dx}{dt} = f(x, u, t) \quad \text{and} \quad y = g(x) \quad (1)$$

where u characterizes the input values, y the output values and x , all other state variables. Thus, the input-output behavior may be expressed as

$$y = \mathfrak{F}_S(u, t) \quad (2)$$

Similarly, the input-output behavior observed in the *in vitro* experiment is expressed as

$$y = \mathfrak{F}'_V(u, t) = \mathfrak{F}_S(u, t) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \quad (3)$$

The degree of similarities between simulated and experimental results can be measured by two functions $E_U(T)$ and $E_T(U)$, as follows

$$\hat{E}_U(T) = \sum_{i=1}^N (Y_i - \mathfrak{F}_S(U, x_i))^2 / N \quad (4)$$

and

$$\hat{E}_T(U) = \sum_{j=1}^N (Y_j - \mathfrak{F}_S(x_j, T))^2 / N \quad (5)$$

$$\text{where } Y_i = \mathfrak{F}'_V(U, x_i) = \mathfrak{F}_S(U, x_i) + \varepsilon \quad (6)$$

and

$$Y_j = \mathfrak{F}'_V(x_j, T) = \mathfrak{F}_S(x_j, T) + \varepsilon \quad (7)$$

Note that

$$E(\hat{E}_U) = \sum_{i=1}^N (\mathfrak{I}_V(U, x_i) - \mathfrak{I}_S(U, x_i))^2 / N + \sigma^2 \quad (8)$$

and

$$E(\hat{E}_T) = \sum_{j=1}^N (\mathfrak{I}_V(x_j, T) - \mathfrak{I}_S(x_j, T))^2 / N + \sigma^2 \quad (9)$$

Also note that, a continuous stochastic representation for E_U and E_T can be given as

$$E_U = \frac{1}{T} \int_0^t (\mathfrak{I}_V(u, t) - \mathfrak{I}_S(u, t))^2 dt + \sigma^2 + \frac{2\sigma}{T} \int_0^t (\mathfrak{I}_V(u, t) - \mathfrak{I}_S(u, t)) dw_t \quad (10)$$

and

$$E_T = \frac{1}{U} \int_0^u (\mathfrak{I}_V(u, t) - \mathfrak{I}_S(u, t))^2 du + \sigma^2 + \frac{2\sigma}{U} \int_0^u (\mathfrak{I}_V(u, t) - \mathfrak{I}_S(u, t)) dw_u \quad (11)$$

If $\mathfrak{I}_V(u, t) = \mathfrak{I}_S(u, t)$, $\forall u$ and $\forall t$, then

$$E(\hat{E}_U) = E(\hat{E}_T) = E_U = E_T = \sigma^2 \quad (12)$$

The mean square error (MSE) was used to measure the goodness-of-fit for comparing the model predictions to the experimental data. In this case, the model yielding lower MSE values was deemed to provide a better fit.

Mathematical modeling and numerical methods

In mammalian caspase-9 pathway, the required components are Apaf-1 (A), caspase-9 (C9), cytochrome c (C), dATP (D), and caspase-3 (C3). The signaling

networks regulate these components as illustrated in Fig.1.1. Based on the traditional caspase-9 pathway model with a specific role for multimeric holoenzyme complex, we developed a kinetic model that was used to represent caspase-9 pathway. The differential equations, variables and default parameters used in this study are summarized in the table 1.3.

Table 1.2 Variable definitions

Abbreviation	Meaning
<i>A</i>	APAF-1 (A_F , Free APAF-1)
<i>D</i>	dATP (D_F , Free dATP)
<i>AD</i>	APAF-1:dATP complex
<i>C</i>	Cytochrome c (C_F , Free cytochrome c)
<i>AC</i>	APAF-1:cytochrome c complex
S_1 (or $(ADC)_1$)	APAF-1:dATP:cytochrome c complex (monomer)
S_2	APAF-1:dATP:cytochrome c complex (dimer)
S_3	APAF-1:dATP:cytochrome c complex (trimer)
S_4	APAF-1:dATP:cytochrome c complex (tetramer)
S_5	APAF-1:dATP:cytochrome c complex (pentamer)
S_6	APAF-1:dATP:cytochrome c complex (hexamer)
S_7	APAF-1:dATP:cytochrome c complex (heptamer)
P_9	Procaspase-9 (P_{9F} , Free procaspase-9)
H_0	Inactive holoenzyme ($S_7:7P_9$ complex)
H_1	Active holoenzyme ($S_7:7C_9$ complex)
C_9	Caspase-9 (C_{9F} , Free and processed caspase-9)
P_3	Procaspase-3 (P_{3F} , Free procaspase-3)
C_3	Caspase-3 (C_{3F} , Free and processed caspase-3)
<i>V</i>	Free DEVD-Afc
<i>F</i>	Free fluorescence molecule Afc

Table 1.3. Differential equations, variable definitions and default parameters

Differential equations for Apaf-1, cytochrome c and dATP

$$\dot{A}_F = -a1 * [A_0][C_0] + a^{-1}[\tilde{A} \bullet C]$$

$$\dot{C}_F = -a1 * [A_0][C_0] + a^{-1}[\tilde{A} \bullet C]$$

$$\dot{D}_F = -b1 * [D_0] + b^{-1}[\tilde{A}C \bullet D]$$

Differential equations for caspase-9

$$\dot{P9}_F = -\delta1 * [S_7][P9_0] + \delta^{-1}[\tilde{S}_7 \bullet P9]$$

$$\dot{C}_9 = k1 * [H_1]$$

Differential equations for caspase-3

$$\dot{P3}_F = -v1 * [P3_0] + v^{-1}[\tilde{S}_3]$$

$$\dot{C}_3 = v2 * [\tilde{S}_3]$$

Differential equations for holoenzyme

$$\dot{H}_0 = +\delta2 * [\tilde{S}_7 \bullet P9_0] - m1 * [H_0]$$

$$\dot{H}_1 = +m1 * [H_0] - k1 * [H_1]$$

Differential equations for DEVD-Afc

$$\dot{V} = -z1 * [V_0] + z^{-1}[\tilde{V}]$$

$$\dot{F} = z2 * [\tilde{V}]$$

Table 1.3. Continued

Variable definitions

A_0	Initial concentration of APAF-1
A_F	Free APAF-1
C_0	Initial concentration of cytochrome c
C_F	Free cytochrome c
D_0	Initial concentration of dATP
D_F	Free dATP
$P9_0$	Initial concentration of procaspase-9
$P9_F$	Free procaspase-9
$P3_0$	Initial concentration of procaspase-3
$P3_F$	Free procaspase-3
H_0	Inactive holoenzyme
H_1	Active holoenzyme
V_0	Initial concentration of DEVD-Afc
F	Fluorescence molecule Afc

Default parameters

$a_1=b_1=1000$, $a^{-1}=b^{-1}=2000$, $a_2=2$, $b_2=0.5$, $\gamma_1=50^*$, $\gamma^{-1}=2000$, $\gamma_2=10$, $\delta_1=140^{**}$, $\delta^{-1}=20000$, $\delta_2=100$, $m_1=3$, $k_1=0.5$, $v_1=6$, $v^{-1}=1000$, $v_2=20$, $z_1=5000$, $z^{-1}=500000$, $z_2=90$ (units, $M^{-1}s^{-1}$); $A_0=4nM$ ((Fearnhead et al., 1998; Zou et al., 1997)), $C_0=1200nM$ (measured by ELISA), $D_0=5000nM$ ((Skoog and Bjursell, 1974)), $P9_0=20nM$ ((Stennicke et al., 1999)), $P3_0=15nM$ ((Stennicke et al., 1998)), $V_0=40000nM$ (added in buffer)

* For the cooperative binding of Apaf-1 complexes: $\gamma_{11}=2$, $\gamma_{21}=4$, $\gamma_{31}=8$, $\gamma_{41}=16$, $\gamma_{51}=32$, $\gamma_{61}=64$ (units, $M^{-1}s^{-1}$).

** For the cooperative binding of procaspase-9: $\delta_{11}=20$, $\delta_{21}=50$, $\delta_{31}=80$, $\delta_{41}=110$, $\delta_{51}=140$, $\delta_{61}=170$, $\delta_{71}=200$ (units, $M^{-1}s^{-1}$).

Modeling APAF-1, dATP and cytochrome c interactions

The first event of intrinsic apoptotic pathway is the release of cytochrome c from mitochondria after exposure to apoptotic inducers. Apaf-1 molecule binds to cytochrome c (C) and acts as the initiator for caspase activation (Cain, 2003). We treat the binding of Apaf-1 and cytochrome c as reversible and the formation of complex as irreversible (Fig.1.2A and C). All abbreviations for the apoptotic components are listed in table 2.

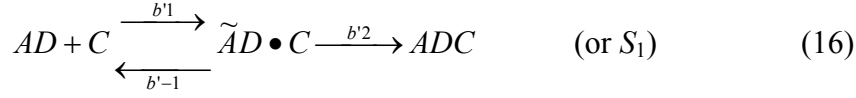
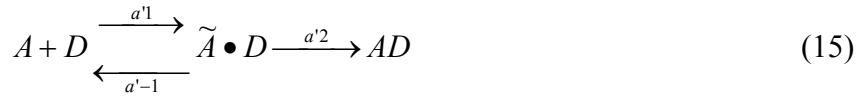


$\tilde{A} \bullet C$ and AC denote the intermediate molecule and Apaf-1:cytochrome c complex, respectively. Previous studies have shown that Apaf-1 also binds and hydrolyzes nucleotides such as dATP. In this reaction, cytochrome c does not affect dATP binding or hydrolysis to Apaf-1 (Zou et al., 1997). Our models treat the binding of dATP and Apaf-1:cytochrome c as reversible and the formation of complex as irreversible (Fig.1.2A and C).



$\tilde{AC} \bullet D$ and ACD denote the intermediate molecule and Apaf-1:cytochrome c:dATP complex (or Apaf-1 monomer), respectively. It is believed that cytochrome c binds to Apaf-1 and dATP stabilizes the binding of cytochrome c to Apaf-1 (Zou et al., 1999). Based on this information, we proposed an alternative model that Apaf-1 molecule

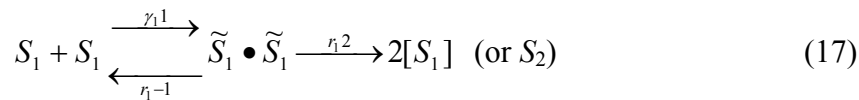
binds to dATP first and acts as the initiator for caspase activation (Fig.1.2B and D).



$\tilde{AD} \bullet C$ and ADC denote the intermediate molecule and Apaf-1:cytochrome c:dATP complex (or Apaf-1 monomer), respectively. We treat the binding of Apaf-1 and dATP as well as Apaf-1:dATP and cytochrome c as reversible and the formation of complex as irreversible.

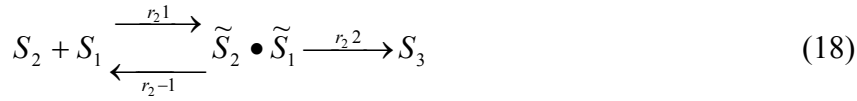
Modeling APAF-1 multimerization

It is known that both dATP and cytochrome c are required for Apaf-1 multimerization. Cytochrome c promotes the multimerization of Apaf-1:cytochrome c complex when the dATP/ATP bound to Apaf-1 is being hydrolyzed (Zou et al., 1999). Previous studies suggest that at least seven Apaf-1 oligomers are involved, as experimentally detected and isolated using gel filtration chromatography (Zou et al., 1999). Furthermore, they also suggest that two Apaf-1 monomers form Apaf-1 dimer complex.

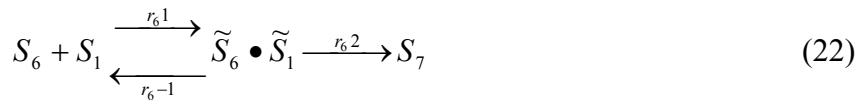
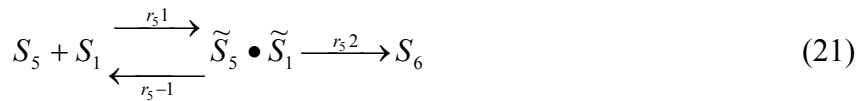
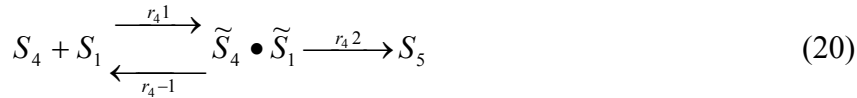
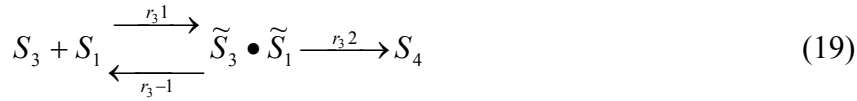


$\tilde{S}_1 \bullet \tilde{S}_1$ and $2[S_1]$ denote the intermediate molecule and Apaf-1:cytochrome c:dATP

dimer complex, respectively. We treat the binding between Apaf-1 complexes as reversible and the formation of complex as irreversible. Apaf-1 dimer complex binds Apaf-1 monomer and forms Apaf-1 trimer complex. Then, Apaf-1 dimer complex and Apaf-1 monomer form Apaf-1 trimer complex.



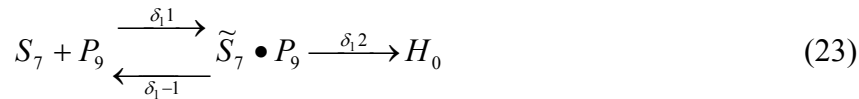
$\tilde{S}_2 \bullet \tilde{S}_1$ and S_2 denote the intermediate molecule and Apaf-1:cytochrome c:dATP trimer complex, respectively. We treat the binding between Apaf-1 dimer complex and Apaf-1 monomer reversible and the formation of complex as irreversible. Then, Apaf-1 trimer complex and Apaf-1 monomer combine to form Apaf-1 tetramer complex. Similar sets of equations hold for Apaf-1 pentamer, hexamer, and heptamer complexes as shown below:



During the multimerization process, if there is a cooperative interaction among Apaf-1 complexes, rate constants of binding between complexes, n th-Apaf-1 complex binding rate, γ_n will continue to increase monotonically ($\gamma_2 \gg \gamma_1$, etc.). However, if there is no cooperative interaction among Apaf-1 complexes, rate constants of binding between complexes, γ_n will remain unaffected ($\gamma_2 = \gamma_1$, etc.).

Modeling the formation and activation of holoenzyme (or Apaf-1:Cytochrome c:dATP:Caspase-9 complex)

Based on previous studies, it is believed that the followings hold: Firstly, Caspase-9 interacts with Apaf-1 only in the presence of dATP and cytochrome c; secondly, Procaspase-9 only binds to the multimeric holoenzyme complex; and finally, Apaf-1 and procaspase-9 are present at an approximately 1:1 ratio in the complex (Acehan et al., 2002; Zou et al., 1999). We made a model for binding between Apaf-1 complex and caspase-9. Apaf-1:cytochrome c:dATP:caspase-9 complex is also known as a holoenzyme.



$\tilde{S}_7 \bullet P_9$ and H_0 denote the intermediate molecule and inactive holoenzyme, respectively. We treat the binding between Apaf-1 heptamer complex and procaspase-

9 reversible and the formation of complex as irreversible. During the binding process, if there is a cooperative interaction among procaspase-9, rate constants of binding between complexes, n th-procaspase-9 binding rate, δ_n1 will be increased ($\delta_21 \gg \delta_11$). However, if there is no cooperative interaction among Apaf-1 complexes, rate constants of binding between complexes, δ_n1 will remain constant ($\delta_21 = \delta_11$). Procaspase-9 is rapidly autoactivated once bound to multimeric Apaf-1 complex.



H_0 and H_1 denote inactive holoenzyme (multimeric Apaf-1 complex with procaspase-9 or $S_7:P_9$) and active form (multimeric Apaf-1 complex with caspase-9 or $S_7:C_9$), respectively. We treat the conversion from inactive holoenzyme to active form as irreversible. Caspase-9 is fully active only when it is bound to Apaf-1 (Rodriguez and Lazebnik, 1999). The activated caspase-9 is released from Apaf-1 complex.



It is assumed that new procaspase-9 has the ability to occupy the empty space in the complex to be processed.

Modeling dual roles of cytochrome c

Once Apaf-1 oligomer is formed, cytochrome c may not be required for next reaction such as caspase-9 binding and activation (Rodriguez and Lazebnik, 1999). However, cytochrome c is still detected in the functional caspase-9-activating complex, which suggests that cytochrome c may have additional role in this pathway (Zou et al., 1999). Therefore, we modeled possible dual roles of cytochrome c in this study. First role is to bind Apaf-1 with dATP consistent with the assumptions of previous studies (eq. (13) and (14)). Then, we assumed that cytochrome c has another possible role, which involves activation of holoenzyme, and incorporated this assumption into the model.



C denotes cytochrome c. We treat the conversion from inactive holoenzyme to active form as irreversible.

Modeling caspase-3 activation

Procaspase-3 is hydrolyzed by active Apaf-1 holoenzyme complex (H_1).



\tilde{S}_3 and C_3 denote intermediate molecule and caspase-3, respectively. We treat the conversion of procaspase-3 into the intermediate molecule as reversible and the formation of active caspase-3, irreversible.

Modeling DEVD-Afc activity

DEVD-Afc is hydrolyzed by free caspase-3.



\tilde{V} and F denote intermediate molecule and fluorescent molecule Afc, respectively. We treat the conversion of DEVD-Afc into the intermediate molecule as reversible and the formation of active caspase-3, irreversible. In this study, we calculate Afc “rate” (fc/min) instead of free “Afc” concentration (nM).

The Robustness of the models

In this study, we have focused on the relative model robustness values of the competing network models to reconstruct the most plausible model, as this approach provides the best metric for the reliability of the computationally simulated models, when compared against small-sample (or non-stationary) experimental data. We have tested the robustness of the models, both extensively and exhaustively, by varying the rate constant values, as most metabolic reactions, as those appearing in apoptosis process, are primarily dependent on k_1/k_{-1} ratio. In our perturbation study, we conducted our parameter-sweep by both increasing and decreasing k_1 values and then simulating the models for each case to evaluate the model robustness.

Part II. Hypothesis Testing with Evolutionary Models

(Drop-out Alignment Allows Homology Recognition and Evolutionary

Analysis of rDNA Intergenic Spacers)

SUMMARY

The rDNA intergenic spacer (IGS) subrepeats play an important role in enhancing RNA Polymerase I transcription, and yet, despite this functional role and presumed selective constraint, they show surprisingly less immediately recognizable conserved similarities than other parts of the rDNA gene. A likely explanation of this paradox could be that the underlying features are simply obscured by the presence of other less relevant repeat-motifs. Support for such a hypothesis comes from two sources: namely, by the observed fast insertion-deletion rates of short mononucleotide micro-satellite runs, which are here referred to as Poly(N) runs (N is any nucleotide), and by their relative abundance in rDNA IGS subrepeats across many species. Some species have different types of IGS subrepeats that do share species-specific Poly(N) run patterns. This finding indicates that many IGS subrepeats within species share a common evolutionary history. Furthermore, by aligning sequences after modifying them by the drop-out method, i.e. by disregarding Poly(N) runs during the sequence aligning step, we sought to uncover evolutionarily shared similarities that fail to be recognized by current alignment programs. To ensure that the improved similarities in the computed alignments are real and not a chance artefact, we calibrated and corrected the IGS subrepeat sequences for the influence of repeat length and estimated the statistical significance of the alignments obtained by the drop-out method by comparing them to null models constructed using random sequence sets from the same genomes. Our analysis led us to conclude that most diverse kinds of rDNA IGS

subrepeats in one species must have been derived from a common ancestral subrepeat, and that it is possible to infer the evolutionary relationships among the IGS subrepeats of different species by comparative genomics methods based on drop-out alignments.

INTRODUCTION

For the last several decades, molecular and evolutionary biologists have been intensely studying the intergenic spacer region (IGS) of the ribosomal RNA genes (rDNA), which separates the 28S and the 18S rDNA coding regions. Not only does the IGS of higher eukaryotes play an important part in RNA polymerase I transcription, but it also contains broadly conserved structural features such as several kinds of repeating elements (or subrepeats), repetitive enhancer elements, duplicated promoters, and conserved secondary structures, which are useful to study in the context of molecular evolution (Baldrige et al., 1992; Kahl, 1988; Reeder, 1989; Ruiz Linares et al., 1991; Sollner-Webb and Tower, 1986). The rDNA IGS, which is composed of the nontranscribed spacer (NTS) and the external transcribed spacer (ETS) regions, contains typically many reiterated subrepeats (Baldrige et al., 1992; Kahl, 1988; Mandal, 1984), with one known exception occurring in *Caenorhabditis elegans*, which has a simple, short structure, and no subrepeats in the IGS region (Ellis et al., 1986). There are length variations of the IGS in most species. However, the IGS has not lent itself as a useful tool for phylogenies of species that are not very closely related, not only because the IGS has large number of reiterated subrepeats but also because the subrepeats' lengths and primary sequences are too different to be aligned (Black et al., 1989; MacIntyre, 1985; Murtif and Rae, 1985; Rogers et al., 1993). Consequently, IGS sequences could be usefully employed only in phylogenetic studies of very closely related species (Bhatia et al., 1996; Borisjuk and Hemleben, 1993;

Cordesse et al., 1993; Da Rocha and Bertrand, 1995; King et al., 1993; Tautz et al., 1987).

In a previous study, we described the rDNA IGS region in the swimming crab, *Charybdis japonica* and reported that the swimming crab IGS also shows a typical IGS structural pattern, which has repetitive subrepeats (Ryu et al., 1999). Especially, three size classes of the swimming crab subrepeats, 60bp, 142bp and 391bp, showed high similarity values, signifying that they shared a common ancestor. We suggested one type of subrepeat (60bp subrepeats, type c) as a prototype for other types. Nevertheless, the primary structures of subrepeats in the swimming crab are quite divergent. One reason for this divergence may be frequent unequal crossing-over and mutation. It has been a well-accepted model that repeated DNA sequences evolve through successive cycles of tandem duplication and divergence of an ancestral sequence (Dover and Tautz, 1986; Grellet et al., 1986; Stark et al., 1989). Similarly, the evolution of the rDNA IGS is thought to include duplication and deletion or divergence processes, resulting in a dynamic change in the subrepeat composition of the IGS (Barker et al., 1988; Cordesse et al., 1993; Ryu et al., 1999). On the other hand, gene conversion and other processes of “concerted evolution” are predicted to maintain similarity among sequences within a subfamily of repeats (Dover and Tautz, 1986). Tandem duplication of single nucleotides, perhaps by polymerase “stuttering”, producing homopolymeric runs, is thought to be another factor resulting in divergence between new types of subrepeats and the original subrepeats (Cunningham et al.,

1991; Jacques et al., 1994). Thus, it is not surprising that DNA sequences have many homopolymeric runs, which are defined as two or more identical consecutive nucleotides. Previous studies showed that genome sequences from many species have long stretches of homopolymeric runs. These homopolymeric runs are also frequently referred to as mononucleotide microsatellites, but will be abbreviated here by the term Poly(N). In general, Poly(A) or Poly(T) runs are more abundant in each taxon than Poly(C) and Poly(G) runs (Toth et al., 2000). Intergenic spacer regions contain more Poly(A/T) over Poly(C/G) in each taxon except *C. elegans* (Toth et al., 2000). But, these distributions differ when constrained to relatively short Poly(N) runs, e.g., 2 to 10bp, and are likely to better address the evolutionary history of rDNA IGS sequences, which contain many short Poly(N) runs than expected.

In this study, we observed fast insertion rates of the Poly(N) in the rDNA IGS and a more rapidly changing rate in rDNA IGS subrepeats. This characterization led us to conclude that changes in Poly(N) could better explain the nature of subrepeat divergence. With the goal of extending this analysis further, we developed a drop-out alignment algorithm, which can mask the differences induced by the Poly(N) runs and recover the obscured underlying phylogenetic signals.

MATERIALS AND METHODS

(1) Sequence Data

The data for various subrepeats of the rDNA IGS region in higher eukaryotes were obtained from published rDNA IGS sequences and GenBank. We used a total of 28 types of available rDNA IGS and 44 types of subrepeats from those species: 3 types of subrepeats from 3 mammalian species, 6 types of subrepeats from 3 amphibian species, 9 types of subrepeats from 5 insect species, 3 types of subrepeats from 2 crustacean species, 22 types of subrepeats from 14 plant species, and one type of subrepeats from one protozoan species (Table 2.1). We also used genomic data from five species, *Xenopus laevis* (National Bioresource Project, version 1.6.5, <http://shigen.lab.nig.ac.jp/xenopus/>), *Drosophila melanogaster* (UCSC genome database, version2, Apr. 2004), human (UCSC genome database, version18, Mar. 2006), mouse (UCSC genome database, version8, Mar. 2006) and rat (UCSC genome database, version4, Nov. 2004).

Table 2.1. List of species and rDNA IGS subrepeats used in this study

	Species	IGS size (GC%:HP%)	Subrepeats (GC%:HP%)	Genbank ID	Ref.
Mammalian	Human (<i>Homo sapiens</i>)	29685(52.1:45.0)*	798(59.8:43.0)*, 282(57.8:45)	U13369	(Gonzalez et al., 1992)
	Rat (<i>Rattus norvegicus</i>)	NTS:2987(47:47)*	150, 572†*	X03838	(Tower et al., 1989; Yavachev et al., 1986)
	Mouse (<i>Mus musculus</i>)	31905(42.1:52.0)*	132*	BK000964	(Kuehn and Arnheim, 1983)
	Hamster (<i>Cricetulus longicaudatus</i>)	NA	356†	M26164†	(Tower et al., 1989)
Amphibian	Tailed frog (<i>Ascaphus truei</i>)	1897(57.2:59.8)*	51(47.1:56.9)S*	X12607	(Morgan and Middleton, 1988)
	Salamander (<i>Triturus vulgaris</i>)	8516(61.6:52.9)*	120(57.5:59.2)*	X98876	(De Lucchini et al., 1997)
	African clawed frog (<i>Xenopus laevis</i>)	NA*	100(82.76)*, 60(74.6:56)*†, 81(76.5: 58)*, 35(75.8:76)*	M23393	(Moss et al., 1980; Pikaard and Reeder, 1988)
	Kenyan clawed frog (<i>Xenopus borealis</i>)	Partial	138	X00184	(Bach et al., 1981)
	<i>Xenopus clivii</i>	Partial	130	V01427	
Lower Cordata	<i>Sea squirt</i> (<i>Herdmania momus</i>)	1880(54.2:36.5)*	NA	X53538	(Degnan et al., 1990)
Insect	<i>Drosophila melanogaster</i>	3632(29.0)*	100, 240†*	AF191295	(Kohorn and Rae, 1982; Ohnishi and Yamamoto, 2004; Simone et al., 1985)
	<i>Drosophila oreana</i>	2385(NA)	241	NA	
	<i>Drosophila hydei</i>	4487(NA)	226	NA	(Murtif and Rae, 1985; Tautz et al., 1987)
	<i>Drosophila virilis</i>	5276(NA)	100(40.0)*, 226(32.0)*	NA	
	<i>Drosophila funebris</i>	4031(30.0)*	NA	L17048	UP
	Bulldog ant (<i>Myrmecia croslandi</i>)	1752(34.3)*	144(31.3)*	AB121789	(Ohnishi and Yamamoto, 2004)
	Asian Tiger Mosquito (<i>Aedes albopictus</i>)	4707 (57.3)*	201(58.2)*, 64(57.8)*, 34(41.2)*, 48(56.3)*	M65063	(Baldrige and Fallon, 1992)
	Yellow fever mosquito (<i>Aedes aegypti</i>)	1797(57.9:42.0)*	49(63.3:35.0)	AF004986	(Wu and Fallon, 1998)
	<i>Bombyx mori</i>	Partial	3, 4, 5, 8	X05086	(Fujiwara and Ishikawa, 1987)
	Tsetse fly (<i>Glossina morisitans</i>)	Partial	420(29.3)*	X05007	(Cross and Dover, 1987)
	Aphid (<i>Acyrtosiphon lactucae</i>)	Partial	247	NA	(Kwon and Ishikawa, 1992)
Crustacean	Swimming crab (<i>Charybdis japonica</i>)	5376(47.6)*	60(45.0)*, 142(48.6),391(43.3)	NA	(Ryu et al., 1999)
	Water flea (<i>Daphnia pulex</i>)	4819 (45.4)*	330(43.3), 200*	U34871	(Crease, 1993)
	Brine shrimp (<i>Artemia cysts</i>)	Partial	618(42.0)*	NA	(Koller et al., 1987)
Fungi	<i>Armillaria jezoensis</i>	611 (43.4)	NA	D89921	DS
	<i>Armillaria ostoyae</i>	590 (43.1)	NA	D89924	(Peyretailade et al., 1998)
	<i>Armillaria sinapina</i>	611 (42.4)	NA	D89925	
	<i>Encephalitozoon cuniculi</i>	1748 (50.8)	NA	AJ005581	

Table 2.1. List of species and rDNA IGS subrepeats used in this study (continued)

	Species	IGS size (GC%:HP%)	Subrepeats (GC%:HP%)	Genbank ID	Ref.
Protozoan	<i>Trypanosoma cruzi</i>	1754	172*	Y00055	(Schnare and Gray, 1982)
Plant	Rice (<i>Oryza sativa</i>)	2140 (71.5)*	254(70.4)*	X54194	(Takaiwa et al., 1990)
	Wheat (<i>Triticum aestivum</i>)	3589(57.0)*	135(61.0)*	X07841	(Barker et al., 1988)
	Oat (<i>Avena sativa</i>)	3982(54.0)*	92(46.0)*, 148(55.0)*	X74820	(Polanco and Perez de la Vega, 1994)
	Fava bean (<i>Vicia faba</i>)	2014(50.0)*	325(50.0)*	X16615	(Kato et al., 1984)
	Hirsuta bean (<i>Vicia hirsute</i>)	2264(48.0)*	379(42.0)*	X62122	
	Adzuki bean (<i>Vigna angularis</i>)	Partial	174(40.0)*	X17210	(Unfried et al., 1991)
	Mung bean (<i>Vigna radiata</i>)	4243(50.0)*	174(43.0)*, 340(57.0)*	X17209	(Unfried et al., 1991)
	Carrot (<i>Daucus carota</i>)	5775(53.0)*	456(56.0)*	D16103	(Suzuki et al., 1996)
	Potato (<i>Solanum tuberosum</i>)	3232(56.5)*	54(63.0)*, 74(57.1)*	X65489	(Borisjuk and Hemleben, 1993)
	Tomato (<i>Lycopersicon esculentum</i>)	3253(53.7)*	63(64.8)*, 141(65.7)*	X14639	(Schmidt-Puchta et al., 1989)
	Cucumber (<i>Cucumis sativus</i>)	3451(61.3)	30(90.0), 90(55.6)*	X07991	(Ganal et al., 1988)
	Squash (<i>Cucurbita maxima</i>)	5508(58.3)*	104(72.4)*, 118(46.1)* 420(47.7), 41(90.1)	X13059	UP
	Strawberry (<i>Fragaria ananassa</i>)	4274(52.0)*	140(54.0)*, 170(52.0)	X58119	UP
	Tobacco (<i>Nicotiana tabacum</i>)	4996(56.0)*	216(68.0), 121(58.0)*	D76443	UP
	<i>Arabidopsis thaliana</i>	4726(49.0)*	495(58.0)*	X15550	(Gruendler et al., 1989)
	Kidney bean (<i>Phaseolus vulgaris</i>)	Partial	166(40.0)*	Z48777	DS
Garden rocket (<i>Eruca sativa</i>)	4003(45.0)*	113(56.0), 120(26.0)*	X74829	(Lakshmikumar and Negi, 1994)	

†: Experimentally tested as an enhancer.

^C: Consensus sequence of the subrepeats

^S: Selected subrepeat with the most similarity value compared to others

*: rDNA IGSs and subrepeats used in this study

UP: Unpublished data

DS: Directly submitted to Genbank

NA: Not available

(2) Random Sequence Testing and Statistical Analysis

To test the accuracy of the drop-out method in subrepeat comparisons, we randomly selected sequences from a full genome database in the same species, and compared them with the subrepeats used in this study. To get a stabilized p value, we repeated the comparisons 10,000 times. All random selections were generated by a random-number-generation function in LISP programming language (Lispworks version 4.2.0, Xanalys Inc). In the random sequence data comparison, the Student's t test was used; a stringent p -value of $p < 0.01$ was considered to be statistically significant.

(3) Sequence Alignment

Sequence alignments were carried out using the global alignment algorithm developed by Needleman and Wunsch (Needleman and Wunsch, 1970) with matching score 2, mismatching score -1, and gap penalty -2. For multiple sequence comparisons, we used the Clustal W (Higgins et al., 1992) method in MegAlign program (version 5.07, DNASTAR Inc) with gap penalty 15, gap length penalty 6. We also used WinDotter (downloaded from the website <http://www.cgb.ki.se/cgb/groups/sonnhammer/Dotter.html>), a dot-matrix program developed by Sonnhammer and Durbin (Sonnhammer and Durbin, 1995) with the default window width of 25 residues and score threshold 35. Some portions of the

alignments were edited by hand to further improve similarity.

RESULTS

2.1 Patterns of Poly(N) in subrepeats of the IGS region

Poly(N) runs up to several nucleotides long appear frequently in the subrepeats of the rDNA IGS region. We hypothesized that this high frequency of the Poly(N) runs was a major factor in the divergence of primary sequence and the length of subrepeats of the rDNA IGS within a species. We collected for analysis 28 rDNA IGS sequences and 44 subrepeats from 28 species including mammals, insects, crustaceans, amphibians and plants. First, we tested for and detected a significant bias in base composition in almost all rDNA IGSs in comparison to genomic sequences (Fig. 2.1A and B). The degrees to which they exhibited these biases in base composition were dramatically increased in almost all subrepeats from both animals and plants (Fig. 2.1C and D). To visualize these biases, we calculated the fractions of the rDNA IGS comprising Poly(N) and plotted them.

For purposes of comparing the observed percentage of Poly(N) in IGS subrepeats against that expected by random clustering of nucleotides, we calculated the expected value of Poly(N) for a given length of random nucleotide sequence as described below. The probability of finding a Poly(N) run of a certain length l depends on the length of the run and the probabilities of having the same nucleotide at either or

both of the adjacent positions.

$$\Pr[\text{Poly}(\text{N})_{i=\text{begin},l,x}] = \Pr[\text{N}_{\text{Prev}}] * \Pr[\text{N}_{\text{Given}}]^l * \Pr[\text{N}_{\text{Next}}] \quad (2-1)$$

where the terms $\Pr[\text{N}_{\text{Prev}}]$, $\Pr[\text{N}_{\text{Given}}]$, $\Pr[\text{N}_{\text{Next}}]$ and l denote, respectively the probability of a different nucleotide at the 5' adjacent site, the probability of a given nucleotide at the site of interest, the probability of a different nucleotide at the 3' adjacent site, and the length of run. Based on the equation 2-1, we can calculate the expected number of Poly(N) in a certain length of sequence. First, the probability (\Pr_{number}) that a Poly(N) of a certain length l and of a certain nucleotide composition x can be calculated by the following mathematical equation if nucleotide composition were random, where $P_x = \Pr[\text{N}_{\text{Given}}]$

$$\begin{aligned} \Pr_{\text{number}} [\text{Poly}(\text{N})_{i=\text{begin},l,x}] \\ = (1 - P_x) * (P_x)^l * (1 - P_x) &= (1 - P_x)^2 * (P_x)^l \end{aligned} \quad (2-2)$$

where i is the first position of a run of nucleotide x of length l . P_x may be estimated from the observed frequency of the nucleotide. The total number of Poly(N_x) runs of nucleotide composition of x and of length 2 or higher is:

$$\begin{aligned} \Pr_{\text{number}} [\text{Poly}(\text{N}_x)_{\text{total}}] \\ = \Pr_{\text{number}} [\text{Poly}(\text{N}_x)_2] + \Pr_{\text{number}} [\text{Poly}(\text{N}_x)_3] + \dots + \Pr_{\text{number}} [\text{Poly}(\text{N}_x)_{m \rightarrow \infty}] \end{aligned} \quad (2-3)$$

This equation can be rewritten using equation 2-2.

$$\begin{aligned}
 \Pr_{\text{number}} [\text{Poly}(\text{N})_{i=\text{begin}, l \geq 2, x}] &= (1 - P_x)^2 * (P_x)^2 * (1 + P_x + P_x^2 + P_x^3 + \dots) \\
 &= (1 - P_x) * (P_x)^2 \\
 \Pr_{\text{number}} [\text{Poly}(\text{N})] &= \sum_{x \text{ in } [A, T, C, G]} (1 - P_x) * (P_x)^2 \\
 &= 4 * \frac{3}{4} * \frac{1}{4} * \frac{1}{4} = 3/16 \qquad (2-4)
 \end{aligned}$$

The last term is obtained under the assumption that all four nucleotides are equiprobable. Thus the expected number of Poly(N)'s in a sequence of length m is $3m/16$.

Second, the expected fraction of the nucleotides in Poly(N) for a certain length of nucleotide sequence also can be calculated if nucleotide composition were random.

$$\begin{aligned}
 \Pr_{\text{fraction}} [\text{Poly}(\text{N})_{i=\text{in}, l, x}] &= (1 - P_x) * (P_x)^l * (1 - P_x) * l \\
 &= (1 - P_x)^2 * (P_x)^l * l \qquad (2-5)
 \end{aligned}$$

where i is in a run of nucleotide x of length l . The total fraction of Poly(N) runs of nucleotides x of length 2 or higher is:

$$\begin{aligned}
 &\Pr_{\text{fraction}} [\text{Poly}(\text{N})_{\text{total}}] \\
 &= \Pr_{\text{fraction}} [\text{Poly}(\text{N})_2] + \Pr_{\text{fraction}} [\text{Poly}(\text{N})_3] + \dots + \Pr_{\text{fraction}} [\text{Poly}(\text{N})_{m \rightarrow \infty}] \qquad (2-6)
 \end{aligned}$$

. This equation can be rewritten by equation 2-5.

$$\begin{aligned}
\text{Pr}_{\text{fraction}} [\text{Poly}(\text{N})_{i=\text{begin}, l \geq 2, x}] &= (1 - P_x)^2 * (P_x) * (2P_x + 3P_x^2 + 4P_x^3 + \dots) \\
&= (1 - P_x)^2 * (P_x) * \left[\left(\frac{2P_x}{1 - P_x} \right) + \left(\frac{P_x^2}{1 - P_x} + \frac{P_x^3}{1 - P_x} + \dots \right) \right] \\
&= (1 - P_x)^2 * (P_x) * \left[\left(\frac{2P_x}{1 - P_x} \right) + \frac{P_x^2}{1 - P_x} (1 + P_x + P_x^2 \dots) \right] \\
&= (1 - P_x)^2 * (P_x) * \left[\left(\frac{2P_x}{1 - P_x} \right) + \left(\frac{P_x^2}{1 - P_x} * \frac{1}{1 - P_x} \right) \right] \\
&= P_x^2 * (2 - P_x) \\
\text{Pr}_{\text{fraction}} [\text{Poly}(\text{N})] &= \sum_{x \text{ in } [\text{A}, \text{T}, \text{C}, \text{G}]} (P_x)^2 * (2 - P_x) = 4 * \frac{1}{4} * \frac{1}{4} * (2 - \frac{1}{4}) \\
&= 7/16 \tag{2-7}
\end{aligned}$$

The last term is obtained under the assumption that all four nucleotides are equiprobable. Thus the expected number of nucleotides in Poly(N)'s in a sequence of length m is $7m/16$. Therefore, the expectation of the content of Poly(N) is 43.75% in any given length of random nucleotide sequence. Also the fraction of sequence that is not in Poly(N) runs is (i.e., a "run" of a single nucleotide):

$$\text{Pr}_{\text{fraction}} [\text{Poly}(\text{N})_{l=1}] = 1 - \text{Pr}_{\text{fraction}} [\text{Poly}(\text{N})_{\text{total}}] \tag{2-8}$$

$$\text{Pr}_{\text{fraction}} [\text{Poly}(\text{N})_{l=1}] = 1 - 7/16 = 9/16 = 56.25\%$$

If we "drop out" the repeated nucleotides except one in each Poly(N) run, the

estimated decrease in the total length of a random sequence with all nucleotides occurring with equal probability of 1/4 is:

$$\begin{aligned} & \Pr_{\text{fraction}} [[\text{Poly}(\text{N})_{i=\text{begin}, l \geq 2, x}] - \Pr_{\text{number}} [[\text{Poly}(\text{N})_{i=\text{begin}, l \geq 2, x}] \\ & = 7/16 - 3/16 = 1/4 \end{aligned} \quad (2-9)$$

Thus, theoretically for a random sequence, its total length will decrease by 25% if all but one of the repeated nucleotides in the Poly(N) runs are dropped out. Using equation 2-7, we may now wish to test whether a certain sequence has unusual patterns of number and fraction of Poly(N) runs: Namely, if the fraction of nucleotides in Poly(N) runs is significantly higher than 43.75%, we may say that the Poly(N) runs appear more frequently than expected in the sequence.

The percentage of the Poly(N) runs in most rDNA IGS subrepeats ranges variously from 17.6% to 76% even though their average ($48.4\% \pm 12\%$ STD) is little bit higher than the expected percentage of Poly(N) runs, 43.75% (Fig. 2.1E). We also investigated the possible relationship between sequence (subrepeat) length and the percentage of Poly(N) runs. The average length of the subrepeats was 148.62bp. Plotting the relationship between Poly(N) percentage and lengths of the subrepeats showed that the different types of subrepeats from the same species have similar Poly(N) percentages, for example, 64bp and 48bp subrepeats from mosquito have

46.9% and 45.8% Poly(N) percentage, respectively.

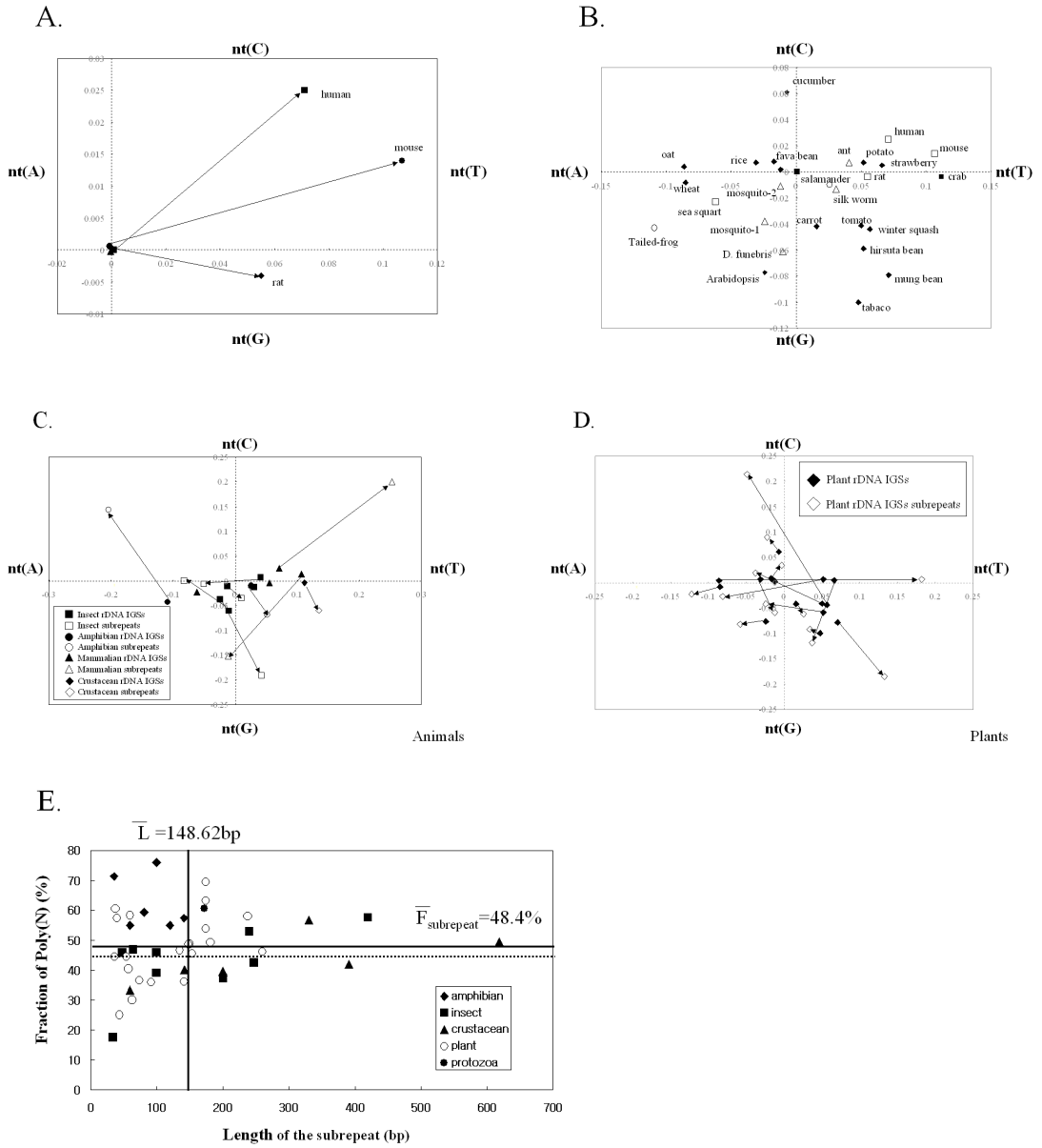


Figure 2.1. Biased base composition of the rDNA IGS and subrepeats. (A) Base composition of mammalian rDNA IGS compared to mammalian genomic sequences. (B) Base composition of full rDNA IGS sequences in many species. Comparison of the base compositions between rDNA IGS and subrepeats in animals (C) and plants (D). (E) Relationship between Poly(N) and lengths of subrepeats. Dashed line indicates the expected content of total Poly(N), which is computed to be 43.75%. Black line indicates the mean values of subrepeat lengths (\bar{L}).

We also analyzed the base patterns of the Poly(N) runs in several species in which rDNA IGS has been well characterized (Fig. 2.2). In this study, we found three different patterns. The 141 (60/81bp) subrepeat of *Xenopus* (Moss et al., 1980) showed the first pattern, or a high frequency of specifically Poly(G/C) runs (Fig. 2.2A). The 100bp subrepeat of *Xenopus* (Moss et al., 1980), the 150bp subrepeat of rice (Takaiwa et al., 1990) and the 64bp and the 201bp subrepeats of mosquito (Baldrige and Fallon, 1992), also have a high frequency of Poly(G/C). Another pattern, a high frequency of Poly(A/T), is apparent in the 420bp subrepeat of the tsetse fly (Cross and Dover, 1987) (Fig. 2.2B). This pattern is also found in the 240bp subrepeat of *D. melanogaster* (Simeone et al., 1985), and the 330bp subrepeat of water flea (Crease, 1993) (data not shown). Thirdly, there is a bias toward a single nucleotide occurring in Poly(N) runs. For example, the 142bp and 390bp subrepeats of the swimming crab (Ryu et al., 1999) have numerous Poly(T)s (Fig. 2.2C); ‘TT’ is found 4 times in the 142bp subrepeats, ‘TTT’ is found about 3 times and ‘TTTT’ as well as ‘TTTTT’ is found just once, whereas ‘AA’ or ‘GG’ is found just once, ‘CC’ is found 10 times and ‘CCC’ is found just once. Therefore, crab subrepeats show primarily the Poly(T) reiterating pattern. Similarly, the 141bp subrepeat of tomato (Schmidt-Puchta et al., 1989) and the 247bp subrepeat of pea aphid show predominantly Poly(G) reiterating patterns (Kwon and Ishikawa, 1992) (data not shown). Lastly, we noted that there are also examples of mixed patterns. The 618bp subrepeat of *Artemia* and the 238bp subrepeat of carrot show minor Poly(A) reiterating patterns as well as reiterating

patterns of all other nucleotides (Koller et al., 1987; Suzuki et al., 1996) (Fig. 2.2D).

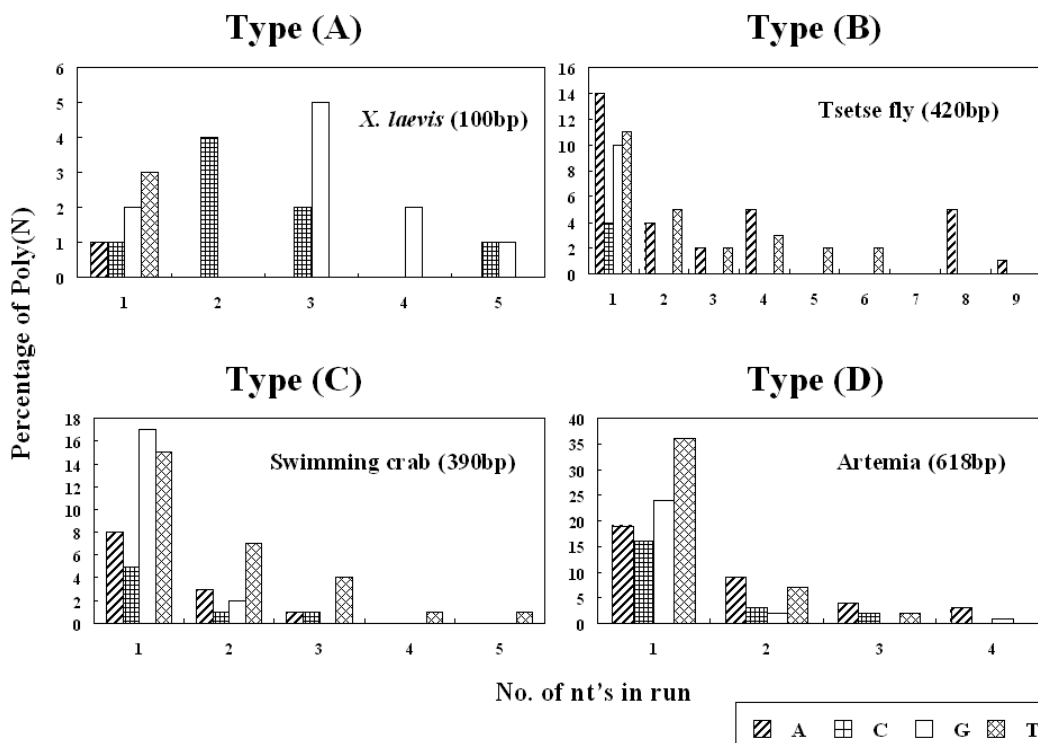


Figure 2.2. The comparison between the Poly(N) base composition of subrepeats changes according to its frequency. Species name and its subrepeat are marked at the top of each graph. The *x*-axis indicates the number of nucleotides repeated and the *y*-axis indicates the frequency. Type A indicates Poly(C and G) reiterating pattern. Type B indicates Poly(A and T) reiteration. Type C indicates single Poly(N) reiterating pattern. Type D indicates mixed patterns.

2.2 Drop-out algorithm

We introduce a novel solution to a sequence alignment problem by using Drop-out method in order to model insertions or deletions of single-letter-runs appearing in long sequences. For comparative genomics applications, these ideas have resulted in a new algorithm that provides efficient alignment over relatively low conserved intergenic spacer regions using different weighting systems and related gap-penalty functions for regular and homopolymeric runs (or Poly(N) for short). Also, for many realistic gap-penalty functions, we show that the drop-out method uses linear space in the worst case, thus avoiding infeasible memory and time burdens. Furthermore, we have evaluated the biological utility of the drop-out method using biological sequences, repeating elements in ribosomal DNA intergenic spacer regions, where drop-out method was capable of discovering many significant correlations than other competing similar alignment tools.

2.2.1 General gap formula

The problem of local pairwise sequence alignment can be described as follows:

For instance, suppose one wishes to compare two sequences X and Y , of sizes m and n , respectively, as shown below:

$X = X_1, X_2, X_3, \dots, X_i, \dots, X_m$

$Y = Y_1, Y_2, Y_3, \dots, Y_j, \dots, Y_n$

We simply assume two functions, a general gap penalty function, w and a score function s . Score function takes two values as input and returns a score based on their match or mismatch. A gap penalty function takes one parameter i , and $w(i)$ denotes the nonnegative penalty given to a gap of length i . We can create a dynamic programming model as follows:

- Let $V(i,j)$ denote the maximum score obtainable for the prefixes $X[1 \dots i]$ and $Y[1 \dots j]$.
- Let $G(i,j)$ denote the maximum score obtainable assuming we align $X[i]$ with $Y[j]$ (regardless of if $X[i]$ and $Y[j]$ match).
- Let $E(i,j)$ denote the maximum score obtainable assuming we align $Y[j]$ against a gap.
- Let $F(i,j)$ denote the maximum score obtainable assuming we align $X[i]$ against a gap.

Score function takes two parameters and produces one value.

$$s(X_i, Y_j) = \begin{cases} 1, & \text{if } X_i = Y_j \\ 0, & \text{if } X_i \neq Y_j \end{cases}$$

In general gap-penalty function, $w(i)$ monotonically increases as i increases, but the rate of increase slows down with increasing values of i ; in other words, $w(i)$ has a

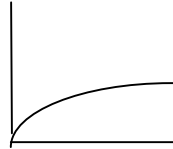
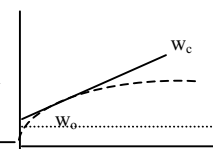
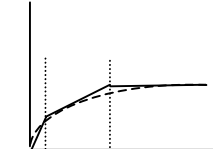
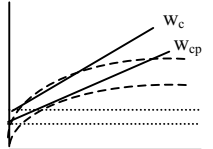
positive derivative, but a negative double-derivative. Therefore, this method discourages long gaps due to high costs. General gap-penalty function uses $O(mn)$ space and $O(mn \times (m+n)) = O(m^2n + mn^2)$ time. Thus, these algorithms are not usable for sequences of length of several kilo-bases.

2.2.2 Affine gap formula

To compare two biological sequences, it is important not to penalize a long gap as the sum of single gaps because one or more insertions/deletions could have happened as one indivisible evolutionary event. Many methods were developed to avoid this problem. Affine gap formula uses low rate slope for long gaps. In Affine gap formula, $w(i)$ is $w_c i + w_o$, where w_c is the rate of increase and w_o is the initial cost. The gap penalty always increases by w_c once the gap is longer than one. Therefore, we need not worry where a gap begins—only whether it already began, or a new gap has just started. In this model, $V(i,0)$ and $V(0,j)$ are equal to $-(w_o + i \cdot w_c)$. Also, $V(i,j)$ and $G(i,j)$ are the same with the general gap penalty formula. For $E(i,j)$ and $F(i,j)$, the gap penalty always increases by w_c once the gap is longer than one. Since this affine gap penalty use $O(mn)$ for both overall runtime and the space, this function is reasonable for using long biological sequences. However, the dynamics of genome evolution shows non-linear rate increasing of the gap penalty. Thus, linear and affine gap penalty function may be unrealistic because it unnaturally forces the algorithm to select smaller gaps every time and thus, to introduce distortions into the underlying

biology.

Table 2-2. The dynamic programming recurrence equation

	General gap	Affine gap	Piecewise-Linear gap (P-part)	Drop-out gap
				
	Length of gap	Length of gap	Length of gap	Length of gap
$V(0,0)$	= 0	= 0	= 0	= 0
$V(i,0)$	= $E(i,0) = -w(i)$	= $E(i,0) = -(w_o + i \cdot w_c)$	= $E_u(i,0) = -w(i)$, if $\exists u(k[u] \leq i < k[u+1])$	= $E(i,0) = -(w_o + i \cdot w_c)$ if gap \neq Poly(N) = $E_p(i,0) = -(w_o + i \cdot w_{cp})$ if gap = Poly(N) (in most case, $w_{cp} \ll w_c$)
$V(0,j)$	= $F(0,j) = -w(j)$	= $F(0,j) = -(w_o + j \cdot w_c)$	= $F_u(0,j) = -w(j)$, if $\exists u(k[u] \leq j < k[u+1])$	= $F(0,j) = -(w_o + j \cdot w_c)$ if gap \neq Poly(N) = $F_p(i,0) = -(w_o + i \cdot w_{cp})$ if gap = Poly(N) (in most case, $w_{cp} \ll w_c$)
$V(i,j)$	$\max\{E(i,j), F(i,j), G(i,j)\}$	$\max\{E(i,j), F(i,j), G(i,j)\}$	$\max\{F_p(i,j), \dots, F_1(i,j), F_0(i,j), E_p(i,j), \dots, E_1(i,j), E_0(i,j), G(i,j)\}$	$\max\{E(i,j), F(i,j), G(i,j)\}$
$G(i,j)$	$V(i-1, j-1) + s(X[i], Y[j])$	$=V(i-1, j-1) + s(X[i], Y[j])$	$V(i-1, j-1) + s(X[i], Y[j])$	$=V(i-1, j-1) + s(X[i], Y[j])$
$E(i,j)$	$\max_{0 \leq k \leq j-1} [V(i, k) - w(j-k)]$	$-w_c + \max\{E(i, j-1), V(i, j-1) - w_o\}$	$E_0(i,j) = -w_c[0] + \max\{E_0(i, j-1), V(i, j-1) - w_o\}$ $E_u(i,j) = \max\{E_u(i, j-1) - w_c[u], E_{u-1}(i, j - (k[u] - k[u-1])) - (k[u] - k[u-1]) \cdot w_c[u-1]\}$, if $j \geq k[u]$; = $-\infty$, otherwise;	$E(i,j) = -w_c + \max\{E(i, j-1), V(i, j-1) - w_o\}$ if gap \neq Poly(N) $E_p(i,j) = -w_{cp} + \max\{E(i, j-1), V(i, j-1) - w_o\}$ if gap = Poly(N)
$F(i,j)$	$\max_{0 \leq k \leq i-1} [V(k, j) - w(i-k)]$	$-w_c + \max\{F(i-1, j), V(i-1, j) - w_o\}$	$F_0(i,j) = -w_c[0] + \max\{F_0(i-1, j), V(i-1, j) - w_o\}$ $F_u(i,j) = \max\{F_u(i-1, j) - w_c[u], F_{u-1}(i - (k[u] - k[u-1]), j) - (k[u] - k[u-1]) \cdot w_c[u-1]\}$, if $i \geq k[u]$; = $-\infty$, otherwise;	$F(i,j) = -w_c + \max\{F(i-1, j), V(i-1, j) - w_o\}$ if gap \neq Poly(N) $F_p(i,j) = -w_{cp} + \max\{E(i-1, j), V(i-1, j) - w_o\}$ if gap = Poly(N)
Space	$O(mn)$	$O(mn)$	$O(mnp) \approx O(mn)$ if p is small	$O(mn)$
Time	$O(mn \times (m+n))$	$O(mn)$	$O(mnp) \approx O(mn)$ if p is small	$O(mnr) \approx O(mn)$ if Poly(N) ratio (r) is small

2.2.3 Piecewise-linear function

Piecewise-linear function is a simple heuristic method to introduce minimal distortions. This method has an ability to decrease the rate of gap penalty by increasing gap penalty so that it is likely to be the one that finds wide usage. Moreover, this method can get the same runtime and space complexity results with simple linear and affine functions. The simplest Piecewise-Linear gap formula is a two-part Piecewise-Linear gap formula. Two-part Piecewise-Linear gap use two extension gap penalties that depend on size k , in addition to the cost of starting a gap w_o ; $w_{c[0]}$ is the rate of penalty increase for any gap below size k , and $w_{c[1]}$ is the rate of increase for any gap of size k or larger.

$$w(i) = \begin{cases} w_{c[0]} \cdot i + w_o, & \text{if } i < k; \\ w_{c[1]} \cdot (i-k) + w_{c[0]} \cdot k + w_o, & \text{if } i \geq k. \end{cases}$$

In two-part Piecewise-Linear gap formula, $G(i,j)$ and $V(i,j)$ behave the same way as before. But, $E(i,j)$ and $F(i,j)$ will have two different parts, 0 and 1. $E_0(i,j)$ denotes the score if we align $Y[j]$ against a gap of length less than k . $E_1(i,j)$ denotes the score if we align $Y[j]$ against a gap of length k or larger. $F_0(i,j)$ and $F_1(i,j)$ behave similarly, but for aligning $X[i]$ against a gap. In this model, for E_0 and F_0 , the gap penalty always increases by $w_{c[0]}$ once the gap is longer than one but smaller than k . Also, for E_1 and F_1 , once the gap is larger than k , the gap penalty always increases by $w_{c[1]}$.

Complex piecewise-linear gap formula is a $(p+1)$ -part piecewise-linear function which

use $(p+1)$ different slopes, denoted $w_{c[0]}$, $w_{c[1]}$, ... and $w_{c[p]}$, as well as p values $k_{[1],k_{[2],\dots,k_{[p]}}$ where the slopes change in addition to a starting gap penalty, w_o . Therefore, this method uses E_0, E_1, \dots, E_p and F_0, F_1, \dots, F_p to compute the scores for aligning $X[i]$ and $Y[j]$ against gaps. The precise reasoning for constructing these tables is similar to the two-part piecewise-linear gap formula. This method use $O(mnp)$ memory and $O(mnp)$ runtime overall. However, in practice for almost all gap penalty functions of interest, p is a small constant. Thus, it is close to $O(mn)$ time and $O(mn)$ space.

2.2.4 Drop-out Gap Formula

Drop-out function leads to a method that can account for insertion of homopolymeric run (or Poly(N) runs) gaps. Consequently this method has the ability to weigh differently, depending on whether the gap involves a Poly(N) or a regular gap (or non-homopolymeric gap). Thus, Drop-out gap function results in a two-parameter algorithm based on different gap penalties for regular insertion and Poly(N) insertion. This method is similar to affine gap formula, $w(i)$ is $w_c i + w_o$, where w_c is the rate of increase and w_o is the initial cost. The gap penalty always increases by w_c once the gap is longer than one. However, w_{cp} will replace w_c if gap is composed of Poly(N).

$$w(i) = \begin{cases} w_o + w_c \cdot i, & \text{if gap} \neq \text{Poly}(N) \\ w_o + w_{cp} \cdot i, & \text{if gap} = \text{Poly}(N) \end{cases}$$

In most case, w_{cp} is much smaller than w_c reflecting an assumed fast insertion rate of Poly(N). Also, $V(i,j)$ and $G(i,j)$ are same as the two earlier dynamic programming tables that were introduced in the algorithm involving the general gap penalty formula. $E_p(i,j)$ denotes the score if we align $Y[j]$ against a gap of Poly(N). $F_p(i,j)$ behaves similarly, but corresponds to the case where $X[i]$ is aligned against a gap.

$$E(i,j) = -w_c + \max \{E(i,j-1), V(i,j-1) - w_o\} \quad \text{if gap} \neq \text{Poly}(N)$$

$$E_p(i,j) = -w_{cp} + \max \{E(i,j-1), V(i,j-1) - w_o\} \quad \text{if gap} = \text{Poly}(N)$$

$$F(i,j) = -w_c + \max \{F(i-1,j), V(i-1,j) - w_o\} \quad \text{if gap} \neq \text{Poly}(N)$$

$$F_p(i,j) = -w_{cp} + \max \{F(i-1,j), V(i-1,j) - w_o\} \quad \text{if gap} = \text{Poly}(N)$$

Since affine gap penalty yields an algorithm with $O(mn)$ complexity for both overall runtime and the space, Drop-out method achieves the same runtime and space complexities with simple linear and affine functions.

2.2.5 One-Piece Drop-out Gap Formula

We can simplify a Drop-out gap formula by using a single value for all Poly(N) gaps, which we shall call One Piece Drop-out Gap Formula. Thus, Poly(N) gaps of all lengths are treated as if they all correspond to a single gap. In this model, the non-Poly(N) gap penalty always increases by w_c once the gap is longer than one. However, the Poly(N) gap penalty has only one value, w_{cp} and it remains a constant regardless of the length of Poly(N) gap.

$$w(i) = \begin{cases} w_o + w_c \cdot i, & \text{if gap} \neq \text{Poly(N)} \\ w_o + w_{cp}, & \text{if gap} = \text{Poly(N)} \end{cases}$$

The simplest way to implement the algorithm is by using modified sequences. During the sequence modification process, Poly(N) runs will be measured and re-expressed as a single nucleotide with a number, q denoting the length of the run. Thus, the smallest number in the sequence will be two as determined by our definition of Poly(N) runs. Suppose now that we wish to compare two sequences X and Y , of sizes m and n , respectively.

$$X = X_1, X_2, X_3, \dots X_i, \dots X_m$$

$$Y = Y_1, Y_2, Y_3, \dots Y_j, \dots Y_n$$

We can rewrite these sequences in the manner shown below, with length m' and n' .

$$X = X_1q_1, X_2q_2, X_3q_3, \dots X_{i'}q_{i'}, \dots X_mq_m'$$

$$Y = Y_1r_1, Y_2r_2, Y_3r_3, \dots Y_{j'}r_{j'}, \dots Y_n'r_n'$$

Thus, if $X_1 = X_2 = X_3$, we write this run of three nucleotides by using the notation X_1q_1 (where $q_1 = 3$)

We illustrate these ideas in the example below:

Original sequences, A and B of respective lengths 23 and 31 are as follows.

Sequence A: 5'- AAAACATATATATGTTTTTCGAT -3'

Sequence B: 5'- ACCCGCGCGCGTGGGGGTCCCCCGGGGAT -3'

Modified sequences with drop-out information: The rewritten sequences A and B are both of equal length 16.

Sequence A: 5'- A4CATATATATGT5CGAT -3'

Sequence B: 5'- AC3GCGCGCGTG5TC5G5AT -3'

One-Piece Drop-out function leads to the same $V(i',j')$ as the affine gap function but produces different $G(i',j')$ with the other gap-penalty functions. Since the sequence is condensed, the Score function will increase by $r_{j'}$ if the matching numbers are more than 2 and $q_{i'}$ is smaller than $r_{j'}$.

$$V(i',j') = \max \{E(i',j'), F(i',j'), G(i',j')\}$$

$$G(i',j') = \begin{cases} V(i'-1,j'-1) + s(X[i'],Y[j']) & \text{if } q_{i'} = r_{j'} = 1 \\ V(i'-1,j'-1) + s(X[i'],Y[j']) \bullet r_{j'} & \text{if } q_{i'} \geq r_{j'} \geq 2 \\ V(i'-1,j'-1) + s(X[i'],Y[j']) \bullet q_{i'} & \text{if } r_{j'} \geq q_{i'} \geq 2 \end{cases}$$

2.2.6 Example of One-Piece Drop-out Gap Function

Below we give an example illustrating the scoring steps of the dynamic programming algorithm with Drop-Out functions. Matching score is 2 and gap penalty (w) is -2.

0	0	A4	C	G	C	G	C	G	C	G	T	G	T5	C	G	A	T
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
C3	0	0	4	2	2	0	2	0	2	0	0	0	0	2	0	0	2
A	0	2	2	4	2	2	0	2	0	2	0	0	0	0	2	2	0
T	0	0	2	2	4	2	2	0	2	0	4	2	2	0	0	2	4
A	0	2	0	2	2	4	2	2	0	2	2	4	2	2	0	2	2
T	0	0	2	0	2	2	4	2	2	0	4	2	6	4	2	0	4
A	0	2	0	2	0	2	2	4	2	2	2	4	4	6	4	4	2
T	0	0	2	0	2	0	2	2	4	2	4	2	6	4	6	4	6
A	0	2	0	2	0	2	0	2	2	4	2	4	4	6	4	8	6
T	0	0	2	0	2	0	2	0	2	2	6	4	6	4	6	6	10
G5	0	0	0	4	2	4	2	4	2	4	4	8	6	6	6	6	8
T	0	0	0	2	4	2	4	2	4	2	6	6	10	8	6	6	8
C5	0	0	2	0	4	4	4	4	4	4	4	6	8	12	10	8	6
G5	0	0	0	4	2	6	4	6	4	6	4	6	6	10	14	12	10
A	0	2	0	2	4	4	6	4	6	4	6	4	6	8	12	16	14
T	0	0	2	0	2	4	4	6	4	6	6	6	6	6	10	14	18

The alignment traced back from the dynamic-programming table, shown above, can be rephrased in terms of the original sequences as follows.

```

Seq A: AAAAC--ATATATATG---TTTTTC---G---AT
      ||      ||      ||      ||      ||      ||
Seq B: A---CCCGCGCGCGTGGGGT---CCCCCGGGGAT
  
```

Sometimes, it may appear somewhat difficult to understand sequence alignments with large portions of Poly(N) gaps. Therefore, we introduced two alignment layout methods. First method is the simplest one with no information of Poly(N) runs. In this

case, alignment can only show a possible relationship between two sequences.

The second method is the layout with information of Poly(N) runs. Therefore, alignment can show a possible relationship between two sequences as well as information about Poly(N) runs.

Layout 1:

```
Seq A: ACATATATATGTCGAT
      ||           |||
Seq B: ACGCGCGCGTGTGAT
```

Layout 2:

```
Seq A: A4C1ATATATATG1T5C1G1AT
      | |           | | | | |
Seq B: A1C3GCGCGCGTG5T1C5G5AT
```

2.2.7 Comparison of One-Piece Drop-out Gap Function with Linear Gap Function

Linear gap function with the same parameters will minimize gaps in order to decrease gap penalty score. Thus long gaps are not favorable in any alignment method with any gap penalty values. For example, Alignment with matching score, 2 and gap penalty, -2 is as follows.

```

Seq A: A---AAACATATATATGTTTTTCG----AT
      |       |       |       ||       ||       ||
Seq B: ACCCGCGCGCGTGGGGGTCCCCCGGGGGAT

```

In this case, we cannot get any information about Poly(N) runs. It is well known that low level of gap penalty also results in a tendency to add gaps in the alignment. Therefore, we can assume that zero or very small gap penalty can create results similar to what one expects with the drop-out alignment. Thus, we ran the same sequence with zero gap penalty.

```

Seq A: AAAA-----CATATATATG----TTTTT-----C-----GAT
      |       |       |       |||      |       |       |||
Seq B: A---CCCGCGCGCG-----TGGGG-----TCCCCCGGGGGAT

```

However, zero gap penalty with the same matching score could not distinguish between Poly(N) gaps (open box with dashed line) and non-Poly(N) (open box with solid line) gaps. Thus, many mismatched nucleotides could be treated as gaps. In this case, alignment will end up having enormous amounts of gap.

2.2.8 Test of drop-out method with biological sequences

We selected two subrepeat sequences from ribosomal DNA (rDNA) intergenic spacer region. *Drosophila virilis* is a prominent reference species for comparison with *Drosophila melanogaster* in regard to patterns and mechanisms of molecular and genomic evolution. We used two subrepeats, 100bp from *D. melanogaster* (DM100) and 100bp from *D. virilis* (DV100). When we used matching score of 2 and a linear gap penalty of -2, we obtained 42.86% identity with 112 informative sites.

Matching score, 2 and gap penalty, -2

```
DM100 -GCCAAACAGCTCGTCATCAATTTAGTG-AC-GCAGG-CATATGATATTG
      | | | | | | | | | | | | | | | | | | | | | | | |
DV100 A-----A--TAGGGAGTGGTGTGCGCGCAAGGCATGTCATATGAAAAAT

DM100 TGTCCCTATCATATAATTAATATAAA-GAATA-TAAAGAATTTTAT--CA
      | | | | | | | | | | | | | | | | | | | | | | | |
DV100 TTTAAATACGGTAA-ATTCATATGACTGTGGACTGTTGACCTTGCCGGC-

DM100 AG-AGT-----A
      | | | | | | |
DV100 ATAAGTCAATTA
```

Identity = 48/112 = 42.86%

In contrast, when we used matching score with 2 and zero gap penalties, we obtained a slightly higher identity (45%) level. However, as we expected, total informative sites were increased to 131 due to many gaps, which are marked above as open boxes with dashed lines. Thus, this alignment is not a proper way to present sequence relationship.

Matching score, 2 and gap penalty, 0

```

DM100  GCCA-A---ACAGCTC-GTCATCAATTTAGTGACGC-AGGCATATGATAT
      | |         || | || | |
DV100  A--ATAGGG--AG-T-GGT-GTC-----G-CGCAAGGCATGTCATA-

DM100  TGTGTCCTATCATATAAT--TA-ATA---TAAA--GAATAT-AAA---G
      | | | | | | | | | | | | | | | | | | | | | | | | | |
DV100  T--G-----A--A-A-AATTTTAAATACGGTAAATTC-ATATGACTGTGG

DM100  AAT-TTTA---T-----CA-AGAGT-----A
      | | | | | | | | | | | | | | | | | | | | | | | |
DV100  ACTGTTGACCTTGCCGGCATA-AGTCAATTA

Identity = 59/131 = 45.0%

```

Next, we used the one-piece drop-out method with the same matching score as the previous method and a gap penalty of -2 . And, we obtained a much higher identity (55.5%) level with fewer informative sites (81). Especially, we found a region with a high similarity score, 82.6% (marked as an open box).

Matching score, 2 and gap penalty, -2

```

DM100d GCACAGCTCGTCATCATAGTGACGCAGCATATGATATGTGTCTATCATAT
      | | | | | | | | | | | | | | | | | | | | | | | | | |
DV100d ATAGAGTGTGTCTCG-CGCAGC-ATGT--CATATGATATACG--TATCATAT

DM100d ATATATAGA-TATAGA-TAT-CAGAGT---A
      | | | | | | | | | | | | | | | | | | | | | | | |
DV100d GACTGT-GACTGT-GACTGCGCATAGTCATA

Identity = 45/81 = 55.5%
Identity of the open box = 19/23 = 82.6%

```

We also used the one-piece drop-out method with 2 matching score and zero gap penalties and obtained similar identity (58.4%) levels.

2.3 The drop-out method and its efficacy in revealing similarity among different types of subrepeats from a species

Because many differences between sequences involve Poly(N) expansions, we reasoned that eliminating Poly(N) runs would reveal similarities that might have been masked by Poly(N) runs. Thus, prior to aligning sequences, we “dropped out” (deleted) all consecutive bases in each Poly(N) run except one base.

The majority of the rDNA IGS region of *Xenopus* is composed of four subrepeats, 35bp, 60bp, 81bp, and 100bp long. The 60bp and 81bp subrepeats are known as enhancer for RNA polymerase I machinery (Pikaard and Reeder, 1988). We found that the percentages of Poly(N) runs in the 35bp subrepeat (74.5%) and the 100bp subrepeat (76%) are higher than the percentage of the 60bp and the 81bp subrepeats, 55% and 59.3%, respectively. When we applied the drop-out method, we found that the three types of subrepeats, 60bp, 81bp and 100bp were more easily aligned, thus revealing possible shared ancestry (Fig. 2.3B). We used Clustal W (Higgins et al., 1992) method in the MegAlign program (version 5.07, DNASTAR Inc) with gap penalty 15, gap length penalty 6. After dropping out the Poly(N) runs, the similarity value between the 35bp and 60bp subrepeats was 41.9% (85.3% increased compared to 22.6% similarity before dropping out Poly(N)s) and the similarity value between the 81bp and 100bp subrepeats was 58.6% (132.5% increase

compared to 25.2% similarity before dropping out Poly(N)s in the Clustal W method with the same condition. We also discovered three conserved regions in the secondary structures: S1 and S2 for the stem region and L1 for the loop region. Both S1 and S2 regions are perfectly complementary (Fig. 2.3C). In order to exercise caution, lest increased similarity might be an artifact resulting from a decrease in overall length due to the drop-out method (as opposed to revealed similarity), we examined the efficiency of the drop-out method using the genomic data. First, we randomly selected the 100bp sequence from the *Xenopus* genome and measured the similarity value with 81bp rDNA IGS subrepeat. Next, we dropped out Poly(N) runs and re-measured similarity value. After that, we compared two similarity values. We repeated this random selection 10,000 times. The average percentage of the Poly(N) before dropping out any Poly(N) runs was $48.8 \pm 0.08\%$ (average \pm standard error); the similarity between a sequence pair increased by about 5.46 ± 0.05 percentage points (average \pm standard error), from $28.08 \pm 0.038\%$ to $33.54 \pm 0.042\%$ (average \pm standard error) after the Poly(N) runs were dropped out. We tested these with the Student *t*-test and found that increased similarity values between two *Xenopus* rDNA subrepeats, 81bp and 100bp were statistically significant ($p < 0.001$). Thus, the increase in similarity owing simply to decrease in length by drop-out alone proved insignificant.

A.

```

X60 -----CAGCCCGACCGGGAGTTCCAGGGA-TCGGGCAGGGGAGCAGGCTCGTCCCC-TGCCCTG-
X81 -----CGGGGACCTGGGGACGGCCCCAGCCCGACCGGGAGTTCCAGGAGCTCGGGCAGAGGGAGCAGGCTCGTCCCCCTGCCCTG--
X100 GAAGAGGGGCCATTCTGAGCCAGGGGACCCGATTTCGGGGTTCGGGGCCCCGGGGGTGCCCGGGGGCCCCGGGGGGG---CGGCTTCCCGGGGTCCCCCGGC
                                     * * * * * * * * * * * * * * * * * * * * * * * * * * * *

```

B.

	S1	L1	S2	
X60d	CAGCGACGAGTCAGA	TCGC	AGAGCA GCTCGTCTGCTG	
X81d	CTGACG CAGCGACGAGTCAGAG	TCGC	AGAGCA GCTCGTCTGCTG	74.0%
X100d	GCA--GACGA-TC-G	TCGC	GTCGTC-GC	58.6%
	** ***** * *	****	* ** ***** **	

C.

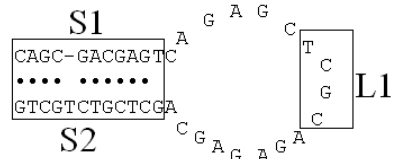


Figure 2.3. Sequence alignment among four subrepeats of the *Xenopus* IGS region. At left margin, four subrepeats are indicated as their lengths. Horizontal lines between bases represent gaps, which have been included for maximum alignment; the vertical bars indicate identical nucleotides in alignment. (A) Multiple alignment among unmodified subrepeat sequences. (B) Subrepeats are modified by the drop-out method which deletes Poly(N) runs except one base in the run. Similarity values (%) are shown at right for some sequence pairs. Asterisks indicate the matched nucleotides between 81bp and 100bp subrepeats. Three conserved regions are marked by S1 (stem 1), S2 (stem 2) and L1 (loop 1). (C) Projected secondary structure of the 60bp subrepeat after dropping out Poly(N). Alignments were generated by Clustal W in MegAlign program (version 5.07, DNASTAR Inc) with gap initiation penalty of 15, and gap extension penalty of 6.

We also tested the dropout method in other species, *Drosophila melanogaster* rDNA IGS subrepeats in order to confirm possible relationship among different types of subrepeats from one species. *D. melanogaster* has two subrepeats, 100bp and 240bp and there exists no reported possible relationship between the two subrepeats, primarily because they fail to align in a statistically significant manner by the existing alignment algorithms. The 240bp subrepeats are known to be enhancer for RNA polymerase I machinery (Kohorn and Rae, 1982). However, these sequences have many Poly(A) and Poly(T) runs that obscure the deeper biological signals. By using the drop-out method, which relieved the alignment ambiguities due to Poly(N)s, we discovered regions with true high similarity between *Drosophila* 100bp and 240bp subrepeats (Fig. 2.4). Conserved regions with 75.9% similarity value are marked by box D in lower panels in Figure 4. To calibrate the significance of this similarity value, we computed a p-value by comparing it with a null model created by randomly selecting unrelated sequences from *D. melanogaster* genomic sequences. We repeated this random selection process 10,000 times by drawing the sequences independently but with replacement. The average Poly(N) percentage of such randomly selected sequences before dropping out any Poly(N) runs was $48.36 \pm 0.086\%$ (average \pm standard error); the similarity between a sequence pair increased by about 1.54 ± 0.023 percentage points (average \pm standard error), from $16.73 \pm 0.018\%$ to $20.24 \pm 0.024\%$ (average \pm standard error) after the Poly(N) runs were dropped out. The Student *t*-test showed that increased similarity values between two *Drosophila* rDNA subrepeats,

100bp and 240bp were statistically much more significant ($p < 0.001$).

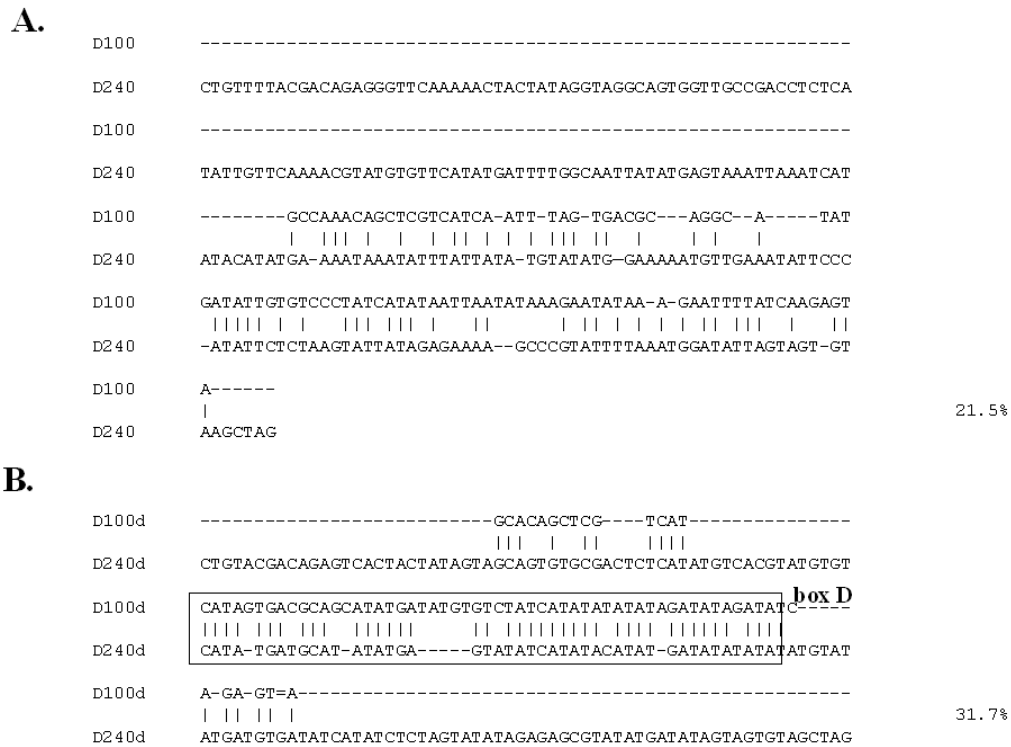


Figure 2.4. Sequence alignment between two subrepeats of *D. melanogaster* IGS region. At left margin, two subrepeats are indicated as their lengths, 100 and 240bp. Upper panel: alignment of unmodified subrepeat sequences. Lower panel: subrepeat sequences modified by the drop-out method aligned with gaps. Conserved region is marked by box D in lower panel. Sequence alignments were carried out using Needleman-Wunsch global alignment algorithm applied with a match score of 2, a mismatch score of -1, and a gap penalty of -2.

In order to ascertain the biological universality of the underlying mechanisms,

we also applied the drop-out method to similar regions in one plant species: namely, the tomato *Lycopersicon esculentum*. The rDNA IGS of tomato is composed of two subrepeat types, 63bp (RE I) and 141bp (RE II) long sequence (Schmidt-Puchta et al., 1989). The Poly(N) percentage of two subrepeat types, 63bp and the 141bp subrepeats of the tomato rDNA IGS is 30% and 36.1%, respectively (Fig. 2.1). In particular, the 141bp subrepeats have a high percentage of guanine, which occurs five times as triplets (GGG) and eight times as dinucleotides (GG). The alignment of “dropped-out” 63bp and 141bp subrepeats resulted in an increase in similarity by 52.7% to 33% compared to 17.4% similarity of the original sequences (Fig. 2.5). Most importantly, tomato 63bp subrepeat showed high similarity value with a 5'-portion of 141bp subrepeat, which we shall refer to as a Box T. The similarity value in this box was 60.6%, much higher number than what would be expected by chance alignment. We also tested the drop-out method in other plant species, potato, *Solanum tuberosum*. Potato rDNA IGS contains two types of subrepeats, or 54bp (type II) and 74bp (type I) subrepeats (Borisjuk and Hemleben, 1993). The similarity value between them using the drop-out method is 60.7% (data not shown), and the analysis uncovered many of the same features as in the other unrelated species.

A.

```

Tom63 -----
Tom141 GGCAGGCGGACGTCATGGGCGTCCTGTGGGCTTAGTAGGCGTGCTGCGTGGGCGCTTGAT

Tom63 -----GGGCAGC-ACACACGGTCGAAC-GACGTC
Tom141          ||||  ||| |  | |  |||||
Tom141 GGCATGCATGGCTCGTCCGTGCTACGCCGTGGGCGTTTACAAAAACATGCAGCGACGTC
Tom63 CGGGCGTGG-CATGCCATCAT-CGCCTTT
Tom141 | | | | | |  | |  |
Tom141 TCGGGGCGAC-TG-----AGGCGG---T

```

43.3%

B.

box T

```

Tom63d GCAGC-ACA-CA--CGTC-G-----A-CGA---CGTCGCG-TGCATGCAT-CATCGC
Tom141d |||| | | | | || |  | |  ||| ||| | | |||| | | | |
Tom141d GCAGCGACGTCATGCGTCTGTGCTAGTAGCGTGCTGCGT-GCGCTG-ATGCATGCAT-GC

Tom63d T-----
Tom141d |
Tom141d TCGTCGTGCTACGCGTGCCTACACATGCAGCGACGTCGCGGACTGAGCGT

```

60.7%

Figure 2.5. Sequence alignment between the 63bp and the 141bp subrepeats of tomato. Upper panel: alignment of unmodified subrepeats sequences. Lower panel: subrepeats modified by the drop-out method that Poly(N) are deleted except one base. The inverted triangle indicates the position of the first nucleotide in bottom panel and corresponding position in upper panel. Highly conserved regions after dropping out are marked by “box T” (62.3% similarity within the box T region). Sequence alignments were carried out using Needleman-Wunsch global alignment algorithm applied with a match score of 2, a mismatch score of -1, and a gap penalty of -2.

2.4 Usage of drop-out alignment method to compare subrepeats between different species

The drop-out method can also be used to compare different subrepeats from different species. First, we tested the drop-out effect on closely related species. We selected several subrepeats from plants because the sizes of plant subrepeats are relatively long and similar to each other. We observed that similarity values were increased from 33.9 ± 0.83 to 56.0 ± 1.98 (mean \pm standard error) by removing Poly(N) runs (Table 2.3). Although mean values are not consistent with evolutionary distance among plant species, we confirmed the usefulness of the drop-out method in finding possible relations among subrepeats. For example, the similarity value of subrepeats between two bean species from the same genus, *Vicia*, was extremely low and unnoticeable (35.8%) before dropping out Poly(N) runs. However, after dropping out Poly(N) runs, the similarity value was found to be relatively high (66.7%) compared with others.

Table 2.3. Comparing similarity values among long rDNA IGS subrepeats from the plant species before and after dropping out Poly(N).

	Arab295	bVF325	bVH379	bVR340	Car456	Rice254
Arab295		36.2 58.1	30.3 54.4	33.3 56.3	31.8* 76.6*	37.9* 49.8*
bVF325			35.8 66.7	32.2 54.5	30.7* 56.7*	27.5 44.5*
bVH379				33.7 53.6	35.4* 56.0*	31.8* 50.2*
bVR340					35.9* 53.4*	37.8* 53.7*
Car456						38.6* 55.7*
Rice254						

* Exclude long gaps generated by size differences at either 5' or 3' end

bVF325: *Vicia faba* 325bp rDNA IGS subrepeat

bVH379: *Vicia hirsuta* 379bp rDNA IGS subrepeat

bVR340: *Vigna radiata* 174bp rDNA IGS subrepeat

Car456: Carrot 456bp rDNA IGS subrepeat

Rice254: Rice 254bp rDNA IGS subrepeat

We also tested the usefulness of the drop-out method in short rDNA IGS subrepeats. The similarity value of subrepeats between two *Drosophila* species, 100bp subrepeats of *D. melanogaster* and *D. virilis* was increased from 36.8% to 58.5% by shrinking Poly(N) runs. Moreover, the similarity value in the high similarity region, BoxD100 (69.4%) was much higher than overall. We observed many species have similar lengths of subrepeats. For example, many species have around 60bp lengths, *Xenopus* (60bp), swimming crab (60bp), tomato (63bp), mosquito (64bp) and potato (54bp). The similarity value obtained from the comparison of *Xenopus* 60bp subrepeat with swimming crab 60bp subrepeat was 46.0%. The 54bp subrepeats of potato and the 63bp subrepeats of tomato also demonstrated a similarity. Fig. 6 represented the drop-out alignment among the 60bp subrepeats of *Xenopus*, the 54bp subrepeat from the potato IGS, and the 63bp subrepeat from the tomato rDNA IGS (Fig. 2.6D). These alignments showed that the similarity between potato 54bp and tomato 63bp subrepeats was increased from 76.8% to 86.4%. The similarity between tomato 63bp and *Xenopus* 60bp subrepeats was increased from 29.7% to 59.6%.

Figure 2.6. Multiple alignments among different subrepeats from various species after dropping out Poly(N)s: Alignment (A) between the 54bp subrepeat of potato and the 63bp subrepeat of tomato (B) between the 60bp subrepeat of *Xenopus* and the 60bp subrepeat of swimming crab (C) between the 100bp subrepeat of *D. melanogaster* and the 100bp subrepeat of *D. virilis* (D) among the 54bp subrepeat of potato, the 63bp subrepeat of tomato and the 60bp subrepeat of *Xenopus*. Vertical bars indicate identical nucleotides. Highly conserved regions in the alignment are marked by a box and labeled in the same manner as in figure 3. At right margin, pair-wise similarities (%) are indicated. Sequence alignments were carried out using Needleman-Wunsch global alignment algorithm applied with a match score of 2, a mismatch score of -1, and a gap penalty of -2.

Discovery of these conserved nucleotides among different subrepeats from the various species, motivated us to explore similar features in many subrepeats from various species. In this alignment, we used 23 types of relatively short subrepeats from 15 species; 11 subrepeats from 7 plant species, 6 subrepeats from 3 insect species, 2 subrepeats from 2 crustacean species, 3 subrepeats from 2 amphibian species and 1 subrepeat from a nematode. The multiple alignments for all these subrepeats are shown in Figure 2.7. In certain cases, to compensate for length variation, we used a part of these subrepeats: either the 5' or the 3' end. We also checked alignments using reverse-complementary sequences to consider a possible gene inversion. We found that most subrepeats shared commonly conserved sequence orders (14 boxes from A1 to A14).



* crusta. = crustacean, amphi. = amphibian, nemat. = nematode

* r = reverse, c = complement, rc = reverse complement, 5' or 3' = 5' or 3' portion of subrepeat

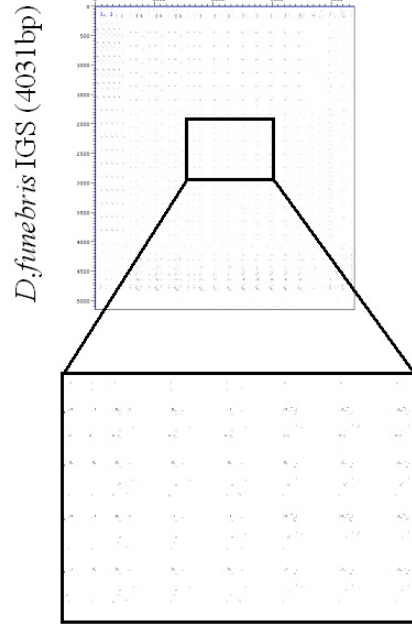
Figure 2.7. Multiple alignments among many subrepeats from various species. Total 23 types of subrepeats from 15 species; 11 subrepeats from 7 plant species, 6 subrepeats from 3 insect species, 2 subrepeats from 2 crustacean species, 3 subrepeats from 2 amphibian species and 1 subrepeats from 1 nematode species. All or portion of subrepeats either 5' or 3' were used. In certain cases, we used have regular, reversed (r), complementary (c) and reverse-complementary (rc) sequences to consider a possible gene inversion or duplication. 5' or 3' indicates 5' or 3' portion of subrepeat. The rectangular boxes marked by A1 to A14 indicate the conserved sequences. Alignments were generated by Clustal W in MegAlign program (version 5.07, DNASTAR Inc) with a gap initiation penalty of 15, and a gap extension penalty of 6.

2.5 Usage of drop-out alignment method to find novel conserved sequences between species

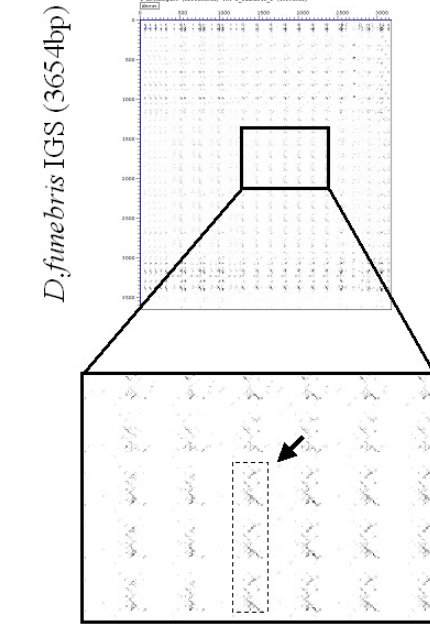
We also applied the drop-out method to whole rDNA IGS sequences. First, we chose two *Drosophila* species, the full rDNA IGS sequences from *D. melanogaster* (4394bp) and *D. funebris* (4031bp), aligned with a Dot-Plot matrix program (Sonnhammer and Durbin, 1995) (Fig. 2.8). The Dot-Plot matrix program is useful to find gene duplications or inversions between sequences. After dropping out Poly(N) runs, the lengths of the rDNA IGS sequences from *D. melanogaster* and *D. funebris* decreased to 3165bp (28% reduction in length) and 3654bp (9.3% reduction in length), respectively. We found that the Dot-Plot matrix revealed many relatively long matched sequences after dropping out Poly(N) (Fig. 2.8B). The clear grid-arrayed diagonals indicate that the rDNA IGS is composed of tandemly reiterated subrepeats that are shared between the species. Although many short diagonal patterns appeared in the dot-plot matrix with original sequences (Fig. 2.8A), they did not extend sufficiently to ascertain detectable similarity between the sequences. We also applied the drop-out method to find matching sequences by using BLAST searching program (window-based BLAST program is available on the NCBI website). We selected 100bp subrepeats from *D. melanogaster* and searched possible matching sequences in full rDNA IGS sequences from *D. funebris*. After dropping out Poly(N) runs, we obtained significantly extended matching sequences (Fig. 2.8D). Based on these

matching sequences, we reconstructed possible 100bp *D. funebris* rDNA IGS subrepeat. The similarity value between 100bp *D. melanogaster* and *D. funebris* rDNA IGS 100bp subrepeats was relatively high, 71.4% (Fig. 2.8E).

A. *D.melanogaster* IGS (4394bp)



B. *D.melanogaster* IGS (3165bp)



C.

Score = 22.3 bits (11), Expect = 0.12
 Identities = 11/11 (100%)
 Strand = Plus / Plus
 Found = 18

Query: 22 ttagtgacgca 32
 |||
 Sbjct: 622 ttagtgacgca 632

D.

Score = 24.3 bits (12), Expect = 0.016
 Identities = 19/20 (95%), Gaps = 1/20 (5%)
 Strand = Plus / Plus
 found = 6

Query: 8 tcgtcatc-atagtgacgca 26
 |||
 Sbjct: 443 tcgtcatctatagtgacgca 462

E.

```

DM100d   GTAGCACAGCTCGTCATC-ATAGTGACGCAGCA-TATGATATGTGTCTATCATATATATATAGATATAGATATCAGA
DF-d     ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
          GTATCATA--TCGTCATCTATAGTGACGCATCACTAT-ATATCTCTATATAGTATAGCGATATATATATGTAGCATG

Sa1-d    TATGACAGTGCAGAGCTCGTCTGCT-CAGTGCA-GAGT-AG
          || ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
X60d     -ATCGCAGAGCAG--CTCGTCTGCTGAGCG-ACGAGTCAG
    
```

73.2%

Figure 2.8. Dot-plot comparisons of whole rDNA IGS sequences from two *Drosophila* species, *D. melanogaster* (4394bp) and *D. funebris* (4031bp). The Figure showed the internal repeat pattern by comparing each sequence with itself in a WinDotter (A dot-matrix program developed by Sonnhammer and Durbin, 1996) with the default window width of 25 residues, score threshold 40 and stringency 10 nucleotides perfect match for all comparisons. Left panel indicates the alignment between two unmodified IGS sequences (A). Right panel indicates the alignment between two IGS sequences (B) after dropping out Poly(N). The small boxes show a magnified window from A and B, respectively. Arrow within the rectangular region indicates grid-arrayed matching sequences between two species. Blast search, Query: 100bp *D. melanogaster* rDNA IGS subrepeats, Target: full rDNA IGS from the *D. funebris* before (C) and after (D) dropping out Poly(N) runs (gap opening:5, gap extension:2). (E) Reconstructed alignment between the 100bp subrepeat of *D. melanogaster* (Dm100d) and possible subrepeat of *D. funebris* (Df-d) (on the top) and between the 60bp subrepeat of *Xenopus* (X60d) and the possible subrepeat of salamander (Sal-d) (on the bottom). Sequence alignments were carried out using Needleman-Wunsch global alignment algorithm applied with a match score of 2, a mismatch score of -1, and a gap penalty of -2.

DISCUSSION

The IGS of many species is known to contain many tandemly reiterating subrepeats. Different copy numbers of repeating elements (or subrepeats) account for most of the length variations of rDNA IGS's among closely related species (King et al., 1993; Tautz et al., 1987). Furthermore, the sequences of the subrepeats themselves differ across species. The analysis described in this paper suggests that most of this variation occurs by a repetition at the nucleotide level, manifested by the occurrence of runs of the same base, or Poly(N) runs. These heterogeneities in size and sequence of subrepeats have made it difficult to compare them directly and to discover common motifs, which may have been conserved during evolution. Nevertheless, the subrepeats may be important in that they are likely to have transcriptional enhancers and promoters for the RNA polymerase I machinery (Labhart and Reeder, 1984; Reeder, 1990).

One of the most important characteristics of rDNA IGS sequences, or occurrence of many short Poly(N)s, is not unique to rDNA IGS subrepeats. In fact, it is a universal character that can be expected theoretically in any given non-coding spacer region or gene as well as in any randomly selected sequence, as estimated by equation 2-1 through 2-3. However, it is important to point out that rDNA IGS subrepeats have certain specific patterns of short Poly(N) runs in some species. These different patterns lead us to believe that these different Poly(N) run patterns might

drive the subrepeats to various evolutionary pathways resulting in diversity both in size and in the primary structure. In comparing Poly(N) percentage and length among various subrepeats, we did not find any specific range of the percentage of Poly(N) for certain type of taxa. Although we found that two amphibian species have unusually higher percentage of Poly(N) when compared to other taxa, we still need more data from other amphibian and related non-amphibian species to conclusively validate this property as a characteristic of amphibians.

Interestingly, we found that many species have their own specific reiterating pattern of Poly(N) runs. The most distinct patterns involve reiterations of guanine/cytosine (G/C) or adenine/thymine (A/T). Moreover, they often share similar frequencies of Poly(G)s and Poly(C)s of same length: for instance, five incidences of CCC would often match with five incidences of GGG. Such patterns could correlate with base pairing in “stems” during the formation of a hypothetical secondary structure or could be a result of the same mutational drive on opposite strands. However, some species show an unequal number of Poly(G/C) or Poly(A/T). Also, some species reiterate predominantly one base. Two hypotheses could explain this phenomenon. First, most polynucleotides in these specific reiterations might be located in loop regions of a presumed secondary structure. Loop regions are more often variable than stem regions. Such variation could be accumulated in regions, which are not subject to strong purifying selection. The Second is adding the same nucleotide by an active mechanism, most likely slippage during replication (Tautz et

al., 1986).

In summary, we devised a drop-out alignment method, and used it to reveal similarities among subrepeats within certain species. We extensively studied the *Xenopus* IGS, which is composed of four subrepeats. Although the 60bp and the 81bp subrepeats are arranged in an alternating pattern and are highly similar to each other (Moss et al., 1980), the other subrepeats, the 35bp and the 100bp, are not (Fig. 2.3A). The *Xenopus* subrepeats were discovered to have an unusually higher percentage of Poly(N). This high Poly(N) content at different locations contributes significantly to the differences between these subrepeats, as is clearly revealed when the drop-out method is applied (Fig. 2.3B) to align them. The similarity value after dropping out the Poly(N) from two *Xenopus* subrepeats, 81bp and 100bp was significantly increased and revealed apparent similarity among all four subrepeat types. We also found high similarity regions between two *Drosophila* subrepeats, 100bp and 240bp after dropping out Poly(N) runs (Fig. 2.4). As another example, the 330bp and 200bp rDNA IGS subrepeats of *Daphnia pulex* are very different from each other not only in their sizes but also in sequence (Crease, 1993) owing to the variability in percentage of the Poly(N). We obtained high similarity value between the two *D. pulex* subrepeats after dropping out Poly(N) runs (data not shown). Similar values of sequence similarities were identified in plant species, tomato and potato rDNA IGS subrepeats.

Perhaps even more dramatically, we discovered that relationships between the

subrepeats of different species could be detected using drop-out alignment. Initially, we focused on sequences around 60bp in length, because most species have small subrepeats around 60bp long, and also because the 60bp subrepeats of *Xenopus* is well known for its function as a transcriptional enhancer (Reeder, 1989). As would be expected, we observed high similarity between the *Xenopus* 60bp subrepeat and the swimming crab 60bp subrepeat as well as between the 54bp subrepeats of potato and the 63bp subrepeats of tomato. Interestingly, the intra-species similarity value between the 81bp and 100bp subrepeats of *Xenopus* was lower than the inter-species similarity between *Xenopus* and swimming crab 60bp subrepeats. Similarly, the degree of identity between the 54bp subrepeats and the 74bp subrepeats of potato is lower than that between the 54bp subrepeats of potato and the 63bp subrepeats of tomato. Considering that there is functional conservation of intergenic spacer elements across distantly related species (Reeder, 1990), and that not all the subrepeats in rDNA IGS play the same role (Robinett et al., 1997), we hypothesized that the subrepeats which have similar functions are more conserved even in distantly related species than other subrepeats with different functions in the same species.

One potential problem with the drop-out method is that it reduces the length of nucleotide sequences by about 25%. This effect could introduce an undesirable bias, since shorter sequences have a better chance to match each other than longer sequences. Using randomized sequences selected from the same genomic sequences, we showed that the longer sequences do display lower similarity values, but not

significantly. Therefore, it is reasonable to conclude that significantly increased similarity values obtained by drop-out alignment reveal otherwise hidden evolutionary homology among rDNA IGS subrepeats.

The drop-out alignment method is also effective in finding conserved sequences in whole IGS sequence comparisons, such as two *Drosophila* rDNA IGS sequences (Fig. 2.8). Although we could not find any significant conservation using unmodified rDNA IGS sequences in a Dot-Plot alignment, application of drop-out method clearly revealed many short matching sequences (Fig. 2.8B). Somewhat serendipitously, we also discovered another possible usage of drop-out method in identifying biologically important regions through a search for matching sequences corresponding to known subrepeats. For *D. funebris*, there exist no reported subrepeats, appearing in previous studies. However, we were able to reconstruct possible rDNA IGS subrepeats in *D. funebris* by using the BLAST search program coupled to the drop-out method. Furthermore, the ability to construct a good multiple alignment of dropped-out IGS subrepeats from different species (Fig. 2.7) also suggests that single-nucleotide Poly(N) runs are the primary reason for the apparent incongruities among these sequences. Thus, we propose that the reiterating nucleotides, resulting in Poly(N) runs, occur at a higher rate than other types of mutations, and thus may have played a greater role in evolutionary changes. This hypothesis is consistent with what has been found through inter-species comparisons of other genes, such as developmental genes in different breeds of dogs (Fondon and Garner, 2004). We believe that the drop-out

method is a useful general tool for searching for deep similarities that may be concealed by distantly or rapidly diverging sequences with fast Poly(N) insertion rates. Other possible applications, for example, would include pretreatment of query and library sequences, using shrinkage of Poly(N)'s, before standard BLAST searches. We are further exploring various generalizations of the drop-out method to reiterating patterns at other levels, such as short reiterating dinucleotides. We thus believe that further studies with the drop-out algorithm and its variants, will provide us with important clues about the evolutionary mechanisms responsible for the diversification of IGS and other sequences.

Appendix A: One –Piece Drop-out Algorithm

(Step 1) The pseudo-code for sequence modification step

```
Make string  $A_m$ 
     $m = \text{length}(\text{sequence } A)$ 
Make string  $B_n$ 
     $n = \text{length}(\text{sequence } B)$ 

count = 0
for  $i=0$  to  $m-1$ 
    if ( $A_{i+1}$  equal  $A_i$ )
        count++
    else    write A and count
           count = 0
for  $j=0$  to  $n-1$ 
    if ( $B_{j+1}$  equal  $B_j$ )
        count++
    else    write B and count
           count = 0

for  $i=0$  to  $m-1$ 
     $F(i,0) = 0$ 
for  $j=0$  to  $n-1$ 
     $F(0,j) = 0$ 
```

(Step 2) The pseudo-code for the scoring step

$w = \text{gap_penalty}$

$S = \text{match_score}$

for $i=1$ to m

 for $j = 1$ to n

 if $((F(m,0) = F(m+1,0))$

 Poly_X++

 if $((F(0,n) = F(0, n+1))$

 Poly_Y++

$a = \text{Poly_X}$

$b = \text{Poly_Y}$

$p = \min(a, b)$

 {

$\text{Diagonal} = F(m-1, n-1) + S * p$

$\text{Down} = F(m-1, n) + w$

$\text{Right} = F(m, n-1) + w$

$F(i,j) = \max(\text{Diagonal}, \text{Down}, \text{Right})$

 }

(Step 3) The pseudo-code for the Tracing back step

```
Alignment_A = ""
Alignment_B = ""
while (i > 0 AND j > 0)
    {
    Diagnol = F(i - 1, j - 1)
    Up = F(i, j - 1)
    Left = F(i - 1, j)

    if ((Diagnol >= Up) and (Diagnol >= Left))
        {
        Alignment_A = add-character (A(i-1))
        Alignment_B = add-character (B(j-1))
        i = i - 1
        j = j - 1
        }

    else if ((Left > Diagnol) and (Left > Up))
        {
        Alignment_A = add-character (A(i-1))
        Alignment_B = add-character ("-")
        i = i - 1
        }

    otherwise ((Up > Diagnol) and (Up > Left))
        {
        Alignment_A = add-character ("-")
        Alignment_B = add-character (B(j-1))
        }
```

```
        j = j - 1
      }
    }
```

(Step 4) The pseudo-code for the finalizing trace

```
while (i > 0)
{
  AlignmentA = add-character (A(i-1))
  AlignmentB = add-character ("-")
  i <- i - 1
}
```

```
while (j > 0)
{
  AlignmentA = add-character ("-")
  AlignmentB = add-character (B(j-1))
  j <- j - 1
}
```

BIBLIOGRAPHY

Acehan D, Jiang X, Morgan DG, Heuser JE, Wang X, Akey CW (2002) Three-dimensional structure of the apoptosome: implications for assembly, procaspase-9 binding, and activation. *Mol Cell* 9:423-32

Antoniotti M, Park F, Policriti A, Ugel N, Mishra B (2003a) Foundations of a query and simulation system for the modeling of biochemical and biological processes. *Pac Symp Biocomput*:116-27

Antoniotti M, Piazza C, Policriti A, Sineoni M, Mishra B (2003b) Modeling cellular behavior with hybrid automata: Bisimulation and collapsing. In (Ed. Priami C). *International workshop on Computational Methods in Systems Biology, LNCS 2602*:57-74

Antoniotti M, Policriti A, Ugel N, Mishra B (2002) XS-systems:Extended S-systems and algebraic differential automata for modeling cellular behaviour. In (Eds. Sahni S, Prasanna VK, Shukla U). *Proceeding of HiPC, LNCS 2552*:431-42

Bach R, Allet B, Crippa M (1981) Sequence organization of the spacer in the ribosomal genes of *Xenopus clivii* and *Xenopus borealis*. *Nucleic Acids Res* 9:5311-30

Baldrige GD, Dalton MW, Fallon AM (1992) Is higher-order structure conserved in eukaryotic ribosomal DNA intergenic spacers? *J Mol Evol* 35:514-23

Baldrige GD, Fallon AM (1992) Primary structure of the ribosomal DNA intergenic spacer from the mosquito, *Aedes albopictus*. *DNA Cell Biol* 11:51-9

Barker RF, Harberd NP, Jarvis MG, Flavell RB (1988) Structure and evolution of the intergenic region in a ribosomal DNA repeat unit of wheat. *J Mol Biol* 201:1-17

Bhatia S, Singh Negi M, Lakshmikumaran M (1996) Structural analysis of the rDNA intergenic spacer of *Brassica nigra*: evolutionary divergence of the spacers of the three diploid *Brassica* species. *J Mol Evol* 43:460-8

- Black WC, McLain DK, Rai KS (1989) Patterns of variation in the rDNA cistron within and among world populations of a mosquito, *Aedes albopictus* (Skuse). *Genetics* 121:539-50
- Borisjuk N, Hemleben V (1993) Nucleotide sequence of the potato rDNA intergenic spacer. *Plant Mol Biol* 21:381-4
- Cain K (2003) Chemical-induced apoptosis: formation of the Apaf-1 apoptosome. *Drug Metab Rev* 35:337-63
- Cain K, Bratton SB, Langlais C, Walker G, Brown DG, Sun XM, Cohen GM (2000) Apaf-1 oligomerizes into biologically active approximately 700-kDa and inactive approximately 1.4-MDa apoptosome complexes. *J Biol Chem* 275:6067-70
- Cordesse F, Cooke R, Tremousaygue D, Grellet F, Delseny M (1993) Fine structure and evolution of the rDNA intergenic spacer in rice and other cereals. *J Mol Evol* 36:369-79
- Crease TJ (1993) Sequence of the intergenic spacer between the 28S and 18S rRNA-encoding genes of the crustacean, *Daphnia pulex*. *Gene* 134:245-9
- Cross NC, Dover GA (1987) Tsetse fly rDNA: an analysis of structure and sequence. *Nucleic Acids Res* 15:15-30
- Cunningham PR, Weitzmann CJ, Ofengand J (1991) SP6 RNA polymerase stutters when initiating from an AAA... sequence. *Nucleic Acids Res* 19:4669-73
- Da Rocha PS, Bertrand H (1995) Structure and comparative analysis of the rDNA intergenic spacer of *Brassica rapa*. Implications for the function and evolution of the Cruciferae spacer. *Eur J Biochem* 229:550-7
- De Lucchini S, Andronico F, Nardi I (1997) Molecular structure of the rDNA intergenic spacer (IGS) in *Triticum*: implications for the hypervariability of rDNA loci. *Chromosoma* 106:315-26

Degnan BM, Yan J, Hawkins CJ, Lavin MF (1990) rRNA genes from the lower chordate *Herdmania momus*: structural similarity with higher eukaryotes. *Nucleic Acids Res* 18:7063-70

Dover GA, Tautz D (1986) Conservation and divergence in multigene families: alternatives to selection and drift. *Philos Trans R Soc Lond B Biol Sci* 312:275-89

Ellis RE, Sulston JE, Coulson AR (1986) The rDNA of *C. elegans*: sequence and structure. *Nucleic Acids Res* 14:2345-64

Evan G, Littlewood T (1998) A matter of life and cell death. *Science* 281:1317-22

Fearnhead HO, Rodriguez J, Govek EE, Guo W, Kobayashi R, Hannon G, Lazebnik YA (1998) Oncogene-dependent apoptosis is mediated by caspase-9. *Proc Natl Acad Sci U S A* 95:13664-9

Fondon JW, 3rd, Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* 101:18058-63

Fujiwara H, Ishikawa H (1987) Structure of the *Bombyx mori* rDNA: initiation site for its transcription. *Nucleic Acids Res* 15:1245-58

Fussenegger M, Bailey JE, Varner J (2000) A mathematical model of caspase function in apoptosis. *Nat Biotechnol* 18:768-74

Ganal M, Torres R, Hemleben V (1988) Complex structure of the ribosomal DNA spacer of *Cucumis sativus* (cucumber). *Mol Gen Genet* 212:548-54

Garcia-Calvo M, Peterson EP, Leiting B, Ruel R, Nicholson DW, Thornberry NA (1998) Inhibition of human caspases by peptide-based and macromolecular inhibitors. *J Biol Chem* 273:32608-13

Gonzalez IL, Wu S, Li WM, Kuo BA, Sylvester JE (1992) Human ribosomal RNA intergenic spacer sequence. *Nucleic Acids Res* 20:5846

Grellet F, Delcasso D, Panabieres F, Delseny M (1986) Organization and evolution of a higher plant alphoid-like satellite DNA sequence. *J Mol Biol* 187:495-507

Gruendler P, Unfried I, Pointner R, Schweizer D (1989) Nucleotide sequence of the 25S-18S ribosomal gene spacer from *Arabidopsis thaliana*. *Nucleic Acids Res* 17:6395-6

Higgins DG, Bleasby AJ, Fuchs R (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput Appl Biosci* 8:189-91

Jacques JP, Hausmann S, Kolakofsky D (1994) Paramyxovirus mRNA editing leads to G deletions as well as insertions. *Embo J* 13:5496-503

Jiang X, Wang X (2000) Cytochrome c promotes caspase-9 activation by inducing nucleotide binding to Apaf-1. *J Biol Chem* 275:31199-203

Jiang X, Wang X (2004) Cytochrome C-mediated apoptosis. *Annu Rev Biochem* 73:87-106

Kahl G (1988) Architecture of eukaryotic genes. VCH Verlagsgesellschaft ; Distribution, USA and Canada, VCH Publishers, Weinheim, Federal Republic of Germany
New York, NY, USA

Kato A, Yakura K, Tanifuji S (1984) Sequence analysis of *Vicia faba* repeated DNA, the FokI repeat element. *Nucleic Acids Res* 12:6415-26

Kim HE, Du F, Fang M, Wang X (2005) Formation of apoptosome is initiated by cytochrome c-induced dATP hydrolysis and subsequent nucleotide exchange on Apaf-1. *Proc Natl Acad Sci U S A* 102:17545-50

King K, Torres RA, Zentgraf U, Hemleben V (1993) Molecular evolution of the intergenic spacer in the nuclear ribosomal RNA genes of cucurbitaceae. *J Mol Evol* 36:144-52

- Kohorn BD, Rae PM (1982) Nontranscribed spacer sequences promote in vitro transcription of *Drosophila* ribosomal DNA. *Nucleic Acids Res* 10:6879-86
- Koller HT, Frondorf KA, Maschner PD, Vaughn JC (1987) In vivo transcription from multiple spacer rRNA gene promoters during early development and evolution of the intergenic spacer in the brine shrimp *Artemia*. *Nucleic Acids Res* 15:5391-411
- Kuehn M, Arnheim N (1983) Nucleotide sequence of the genetically labile repeated elements 5' to the origin of mouse rRNA transcription. *Nucleic Acids Res* 11:211-24
- Kwon OY, Ishikawa H (1992) Unique structure in the intergenic and 5' external transcribed spacer of the ribosomal RNA gene from the pea aphid *Acyrtosiphon pisum*. *Eur J Biochem* 206:935-40
- Labhart P, Reeder RH (1984) Enhancer-like properties of the 60/81 bp elements in the ribosomal gene spacer of *Xenopus laevis*. *Cell* 37:285-9
- Lakshmikumaran M, Negi MS (1994) Structural analysis of two length variants of the rDNA intergenic spacer from *Eruca sativa*. *Plant Mol Biol* 24:915-27
- Legewie S, Bluthgen N, Herzel H (2006) Mathematical Modeling Identifies Inhibitors of Apoptosis as Mediators of Positive Feedback and Bistability. *PLoS Comput Biol* 2
- MacIntyre RJ (1985) *Molecular evolutionary genetics*. Plenum Press, New York
- Mandal RK (1984) The organization and transcription of eukaryotic ribosomal RNA genes. *Prog Nucleic Acid Res Mol Biol* 31:115-60
- Mishra B, Antoniotti M, Paxia S, Ugel N (2005) Simpathica: A computational systems biology tool within the valis Bioinformatics environment. In: Eiles E, Kriete A (eds) *Computational Systems Biology*
- Moretti A, Weig HJ, Ott T, Seyfarth M, Holthoff HP, Grewe D, Gillitzer A, Bott-Flugel L, Schomig A, Ungerer M, Laugwitz KL (2002) Essential myosin light chain as a target for caspase-3 in failing myocardium. *Proc Natl Acad Sci U S A* 99:11860-5

Morgan GT, Middleton KM (1988) Organization and sequence of the compact rDNA spacer of the tailed frog, *Ascaphus truei*. *Nucleic Acids Res* 16:10917

Moss T, Boseley PG, Birnstiel ML (1980) More ribosomal spacer sequences from *Xenopus laevis*. *Nucleic Acids Res* 8:467-85

Murtif VL, Rae PM (1985) In vivo transcription of rDNA spacers in *Drosophila*. *Nucleic Acids Res* 13:3221-39

Nakabayashi J, Sasaki A (2006) A mathematical model for apoptosome assembly: the optimal cytochrome c/Apaf-1 ratio. *J Theor Biol* 242:280-7

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443-53

Ohnishi H, Yamamoto MT (2004) The structure of a single unit of ribosomal RNA gene (rDNA) including intergenic subrepeats in the Australian bulldog ant *Myrmecia croslandi* (Hymenoptera: Formicidae). *Zoolog Sci* 21:139-46

Paxia S, Rudra A, Zhou Y, Mishra B (2002) A random walk down the genomes: DNA evolution in VALIS. *Computer*:73-79

Peyretailade E, Biderre C, Peyret P, Duffieux F, Metenier G, Gouy M, Michot B, Vivares CP (1998) Microsporidian *Encephalitozoon cuniculi*, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. *Nucleic Acids Res* 26:3513-20

Pikaard CS, Reeder RH (1988) Sequence elements essential for function of the *Xenopus laevis* ribosomal DNA enhancers. *Mol Cell Biol* 8:4282-8

Polanco C, Perez de la Vega M (1994) The structure of the rDNA intergenic spacer of *Avena sativa* L.: a comparative study. *Plant Mol Biol* 25:751-6

Purring C, Zou H, Wang X, McLendon G (1999) Stoichiometry, Free Energy and Kinetic Aspects of Cytochrome c: Apaf-1 Binding in Apoptosis. *J. Am. Chem. Soc.*

121:7435-7436

Purring-Koch C, McLendon G (2000) Cytochrome c binding to Apaf-1: the effects of dATP and ionic strength. *Proc Natl Acad Sci U S A* 97:11928-31

Reeder RH (1989) Regulatory elements of the generic ribosomal gene. *Curr Opin Cell Biol* 1:466-74

Reeder RH (1990) rRNA synthesis in the nucleolus. *Trends Genet* 6:390-5

Riedl SJ, Shi Y (2004) Molecular mechanisms of caspase regulation during apoptosis. *Nat Rev Mol Cell Biol* 5:897-907

Robinett CC, O'Connor A, Dunaway M (1997) The repeat organizer, a specialized insulator element within the intergenic spacer of the *Xenopus* rRNA genes. *Mol Cell Biol* 17:2866-75

Rodriguez J, Lazebnik Y (1999) Caspase-9 and APAF-1 form an active holoenzyme. *Genes Dev* 13:3179-84

Rogers SO, Beaulieu GC, Bendich AJ (1993) Comparative studies of gene copy number. *Methods Enzymol* 224:243-51

Ruiz Linares A, Hancock JM, Dover GA (1991) Secondary structure constraints on the evolution of *Drosophila* 28 S ribosomal RNA expansion segments. *J Mol Biol* 219:381-90

Ryan CA, Stennicke HR, Nava VE, Burch JB, Hardwick JM, Salvesen GS (2002) Inhibitor specificity of recombinant and endogenous caspase-9. *Biochem J* 366:595-601

Ryu SH, Do YK, Hwang UW, Choe CP, Kim W (1999) Ribosomal DNA intergenic spacer of the swimming crab, *Charybdis japonica*. *J Mol Evol* 49:806-9

Schmidt-Puchta W, Gunther I, Sanger HL (1989) Nucleotide sequence of the

intergenic spacer (IGS) of the tomato ribosomal DNA. *Plant Mol Biol* 13:251-3

Schnare MN, Gray MW (1982) Nucleotide sequence of an exceptionally long 5.8S ribosomal RNA from *Crithidia fasciculata*. *Nucleic Acids Res* 10:2085-92

Simeone A, La Volpe A, Boncinelli E (1985) Nucleotide sequence of a complete ribosomal spacer of *D. melanogaster*. *Nucleic Acids Res* 13:1089-101

Skoog L, Bjursell G (1974) Nuclear and cytoplasmic pools of deoxyribonucleoside triphosphates in Chinese hamster ovary cells. *J Biol Chem* 249:6434-8

Sollner-Webb B, Tower J (1986) Transcription of cloned eukaryotic ribosomal RNA genes. *Annu Rev Biochem* 55:801-30

Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167:GC1-10

Stark GR, Debatisse M, Giulotto E, Wahl GM (1989) Recent progress in understanding mechanisms of mammalian DNA amplification. *Cell* 57:901-8

Stennicke HR, Deveraux QL, Humke EW, Reed JC, Dixit VM, Salvesen GS (1999) Caspase-9 can be activated without proteolytic processing. *J Biol Chem* 274:8359-62

Stennicke HR, Jurgensmeier JM, Shin H, Deveraux Q, Wolf BB, Yang X, Zhou Q, Ellerby HM, Ellerby LM, Bredesen D, Green DR, Reed JC, Froelich CJ, Salvesen GS (1998) Pro-caspase-3 is a major physiologic target of caspase-8. *J Biol Chem* 273:27084-90

Stucki JW, Simon HU (2005) Mathematical modeling of the regulation of caspase-3 activation and degradation. *J Theor Biol* 234:123-31

Suzuki A, Tanifuji S, Komeda Y, Kato A (1996) Structural and functional characterization of the intergenic spacer region of the rDNA in *Daucus carota*. *Plant Cell Physiol* 37:233-8

- Takaiwa F, Kikuchi S, Oono K (1990) The complete nucleotide sequence of the intergenic spacer between 25S and 17S rDNAs in rice. *Plant Mol Biol* 15:933-5
- Tautz D, Tautz C, Webb D, Dover GA (1987) Evolutionary divergence of promoters and spacers in the rDNA family of four *Drosophila* species. Implications for molecular coevolution in multigene families. *J Mol Biol* 195:525-42
- Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 322:652-6
- Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 10:967-81
- Tower J, Henderson SL, Dougherty KM, Wejksnora PJ, Sollner-Webb B (1989) An RNA polymerase I promoter located in the CHO and mouse ribosomal DNA spacers: functional analysis and factor and sequence requirements. *Mol Cell Biol* 9:1513-25
- Unfried K, Schiebel K, Hemleben V (1991) Subrepeats of rDNA intergenic spacer present as prominent independent satellite DNA in *Vigna radiata* but not in *Vigna angularis*. *Gene* 99:63-8
- Voit EO (1991) Canonical nonlinear modeling : S-system approach to understanding complexity. Van Nostrand Reinhold, New York
- Voit EO (2000) Computational analysis of biochemical systems : a practical guide for biochemists and molecular biologists. Cambridge University Press, Cambridge ; New York
- Wu CC, Fallon AM (1998) Analysis of a ribosomal DNA intergenic spacer region from the yellow fever mosquito, *Aedes aegypti*. *Insect Mol Biol* 7:19-29
- Yavachev LP, Georgiev OI, Braga EA, Avdonina TA, Bogomolova AE, Zhurkin VB, Nosikov VV, Hadjiolov AA (1986) Nucleotide sequence analysis of the spacer regions flanking the rat rRNA transcription unit and identification of repetitive elements. *Nucleic Acids Res* 14:2799-810

Yu X, Acehan D, Menetret JF, Booth CR, Ludtke SJ, Riedl SJ, Shi Y, Wang X, Akey CW (2005) A structure of the human apoptosome at 12.8 Å resolution provides insights into this cell death platform. *Structure* 13:1725-35

Zou H, Henzel WJ, Liu X, Lutschg A, Wang X (1997) Apaf-1, a human protein homologous to *C. elegans* CED-4, participates in cytochrome c-dependent activation of caspase-3. *Cell* 90:405-13

Zou H, Li Y, Liu X, Wang X (1999) An APAF-1-cytochrome c multimeric complex is a functional apoptosome that activates procaspase-9. *J Biol Chem* 274:11549-56

Zubay G (1993) *Biochemistry*. Wm. C. Brown. Publishers