MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Bldg. 68-217, Cambridge, MA 02139-4307

Oct 24, 2005

Yves Brun
Systems Biology/Microbiology Faculty Search
Department of Biology
Indiana University
Jordan Hall 142
Bloomington IN 47405-7005

Dear Members of Search Committee,

I am writing to apply for the open faculty position in your department. I am currently a postdoctoral fellow in Dr. Chris Burge's lab at MIT, where I have used a combination of experimental and computational approaches to study the amazing biology of RNA.

I received my doctorate from Johns Hopkins School of Medicine in the lab of Dr. Paul Englund. During my thesis research, I studied RNA interference in *Trypanosoma brucei*, a parasitic protozoa that causes human diseases.

With an enthusiasm for RNA biology, I joined the computational biology lab of Dr. Chris Burge, who studies regulation of gene expression at the RNA level. Here I set up, for the first time, the experimental part of this lab. I also learned computational biology from one of the best people in this field. My first project was to design an *in vivo* method to systematically screen for exonic splicing silencers (ESSs), a class of understudied *cis*-regulatory elements that often contribute to alternative splicing. The roles of ESSs were further examined through extensive bioinformatic analysis to yield a new hypothesis of their function, which was further tested experimentally. This project represents a high degree of synergy between experimental and computational approaches, and provides a better understanding of how splicing is regulated in systematic level (see my research accomplishment for details).

Pre-mRNA splicing, and alternative splicing in particular, has emerged as a key regulatory step of gene expression in humans. In the future, I plan to systematically investigate the information that regulates splicing (see my research plan for more details). In addition to pursuing my own research, I believe I could benefit others in the department, especially those who are interested in RNA or computational biology. I am also eager to teach scientists of next generation, and I expect mutual inspirations with the students.

Through the course of my graduate and post-doc study, I have gained experience in writing research papers and reviews. In addition, I have successfully written for a post-doctoral grant and assisted in writing several R01 grants.

Letters of recommendation will be mailed separately by Drs. Chris B. Burge, Paul T. Englund, and Phillip A. Sharp. Thank you for considering my application.

Sincerely

Zefeng Wang

Enclosed: (cv, research statement and teaching statement)

Research Accomplishments and Future Plans

A major surprise from human genome is that the total number of protein coding genes is much lower than previously estimated (the latest prediction is 20,000 to 25,000). This finding indicated that additional genomic complexity might be added at the level of RNA processing. More than 60% of human genes undergo alternative splicing, and the disruption of splicing by mutations can cause various genetic diseases (e.g. 175 splicing mutations are found in current Cystic Fibrosis Mutation Database). The information governing alternative splicing is just beginning to be understood, and my major research interest is to thoroughly investigate such information using system biology approaches. Specifically, I will (i) systematically identify and characterize the sequence elements and protein factors that regulate splicing; (ii) integrate such knowledge to understand how splicing is regulated, and eventually to predict the splicing decisions of any transcript in different cell types; (iii) develop a high throughput platform to screen for chemicals that modulate splicing.

Research accomplishments:

1. RNA interference in trypanosomes

My experience with RNA biology started in my dissertation research with Dr. Paul Englund studying *Trypanosoma brucei*, a protozoan parasite that causes devastating diseases in Africa. In 1999, I was introduced to RNA interference at a Woods Hole summer course. Back then, the genetic manipulation of trypanosomes had been cumbersome, and the power of RNAi as a genetic tool was just starting to be explored. I developed an inducible RNAi system by expressing dsRNA from a stably integrated vector. This approach had allowed me to gain insight into the parasite's biology, in particular that concerning the replication of trypanosome's unique mitochondrial DNA (termed kinetoplast DNA or kDNA, a network of thousands of DNA circles that are topologically interlocked in a planar array). Using RNAi, I successfully determined the function of a DNA topoisomerase II in kDNA replication, and further clarified how the trypanosome maintains kDNA network size. This work was published in two *EMBO Journal* papers.

To further explore the power of RNAi as genetic tools, I collaborated with James Morris and Mark Drew to develop an RNAi vector (called pZJM) to simplify the induction of dsRNA in trypanosomes. This vector has since then become a standard reagent to silence genes in trypanosomes, and was distributed to more than 50 labs around the world. We also developed an RNAi based genetic screen with pZJM containing a library of genomic DNA fragments. This RNAi library, the first in any organism, allowed us to do forward genetic screen to find new gene(s) responsible for the biological processes of interest. This work was described in three papers in *EMBO Journal* and *JBC*, and James Morris, Mark Drew and I were co-first authors on all of them.

2. Systematic study of cis-elements that regulate splicing

My enthusiasm for RNA biology and for studies on a genomic scale led me to a postdoctoral position in Dr. Chris Burge's lab. Here I turned to the question of how splicing specificity is determined. It had been shown in human transcripts that splice sites contain only part of the information required for accurate splicing (Lim and Burge, 2001 PNAS 98: 1193-8). Further, human transcripts contain numerous 'decoy' splice sites that are almost never used. Despite this potential for sloppiness, the splicing occurs with remarkable precision, suggesting that additional information must contribute to determining the splicing specificity. The prime candidates are *cis*-elements in exons or introns that either enhance or silence the usage of adjacent splice sites. Depending on their locations and effects on splicing, these elements are defined as exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs).

ESSs are exonic *cis*-regulatory elements that inhibit the use of adjacent splice sites, often contributing to alternative splicing. Although very abundant in the human genome, only a handful of ESSs had been identified by mutational analysis. Most of these examples share little similarity, suggesting that many more remain to be discovered. To systematically identify ESSs, I developed a cell-based splicing reporter system to screen a library of random decamers for ESS (Fig. 1). Our screen, which we call the Fluorescence-Activated Screen for Exonic Splicing Silencers (FAS-ESS, or FAS for short), has identified more than one hundred ESS decamers. Most ESSs can be clustered into groups to yield seven putative ESS motifs. Potential roles of ESSs in splicing were explored using various computational methods including the development of ExonScan, an algorithm that simulates splicing based on known or putative splicing-related motifs. These computational analyses shed light on how ESSs regulate splicing *in vivo*. This work combined the strength of a large-scale screen and computational analysis, and a similar strategy can be used to systematically identify and analyze other splicing regulatory elements (see Future plan 1). A paper describing this work is published in *Cell*.

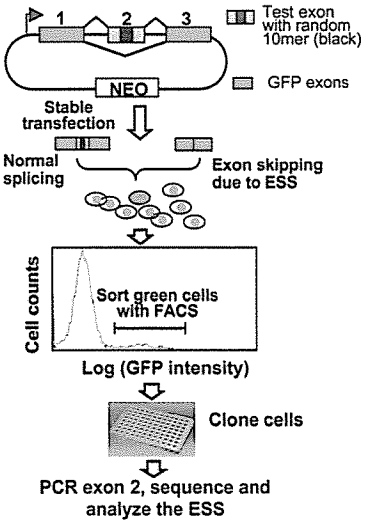

Figure 1. Scheme of FAS-ESS. Exon 1 and 3 of the mini-gene reporter form a complete mRNA of GFP, and exon 2 was inserted with random 10mers. Cells with ESS in exon 2 will skip exon 2 to produce functional GFP, which can be detected and sorted with FACS. The ESS insertion can subsequently be identified by PCR/sequencing.

I have further adopted the FAS strategy to screen for ISS by inserting random sequences into the intronic region close to splice sites rather than into the exon 2 (Fig. 1). The rest of the screen and analysis will be similar to those of ESS. To date more than a hundred ISSs decamers were identified from the screen, and they were clustered into different groups to yield ISS motifs. The roles of FAS-ISS in splicing regulation are being analyzed with various computational and experimental approaches. This work represents the first *in vivo* identification of ISS in large scale, and we are preparing a manuscript to describe this.

3. Splice sites definition by ESS

A key question of splicing is how the splicing machinery can choose the authentic splice site among all the decoy sites. This is especially important to the alternative splicing process in which certain splice site has to be chosen among the alternative sites. Our analyses of FAS-ESS suggested a role of ESS in choosing real splice sites over the decoy sites. This new role, which we called splice site definition, provided a possible mechanism to regulate the selection of alternatively spliced sites. I tested this hypothesis with reporter systems containing dual splice sites at either end of an exon. Sequences representing all major FAS-ESS groups were inserted between the two splice sites, and these ESSs promote the use of the distal sites in all cases. ESSs were found to be enriched in the exonic regions between alternative sites, and mutation of such ESS favors the use of proximal sites, suggesting a general role for ESSs to regulate splice site choice. ESSs were also found to be conserved between two alternatively spliced sites. Recruiting of the *trans*-factors binding to ESS can recapitulate the splice site definition activities. Additional experiments have inferred the mechanism by which ESSs help to define splice site. A manuscript describing this will be submitted soon to *Science*.

Ongoing Work:

Identify small RNA in filamentous fungi

I am working on another project that involves identification of small RNA that may participate in gene regulation in filamentous fungi. Collaborating with Dr. David Bartel's lab, I have cloned about 1000 small RNA with a length of 20 to 30 nt. Further characterization of the structures and functions of these small RNAs is underway.

Future Plans:

The main goal of this proposal is to systematically analyze the information that regulates splicing using a combination of experimental and computational tools. I will pursue this goal by identifying and analyzing *cis*-elements and protein factors that regulate splicing, and screening for compounds that modulate the splicing process.

1. Systematically identify and characterize the factors that regulate splicing

- ***Characterization and analysis of additional cis-elements that regulate splicing***

Among the splicing regulatory *cis*-elements, intronic elements are relatively understudied. However, they could be equally important in splicing regulation.

The success of the FAS approach has laid the foundation for identifying additional *cis*-elements. I will modify this approach to screen for ISE. An intron with "weak" splice sites will be used to separate the two GFP exons, thus the correct splicing of GFP gene should depend on the presence of ISE in the "weak" intron. Using very similar strategies (Fig. 1), I will screen a random sequence library for ISEs that enhance the splicing of weak sites to produce functional GFP. Because there is limited knowledge about ISE, the screen and subsequent analyses are expected to provide insight into how the splicing specificity is defined.

Systematic identification of all the *cis*-elements will pave the way for further analysis of the roles of these sequences by various computational methods, including a splicing simulator like ExonScan. The correlation between *cis*-regulatory elements and the process of alternative splicing will also be studied computationally (e.g. using ExonScan or comparative genomics) and experimentally (e.g. by mutagenesis). In addition, many disease mutations can cause abnormal splicing by affecting the *cis*-regulatory elements. After cataloging both the *cis*-elements and mutations, I will fully examine how these splicing mutations affect the *cis*-elements.

Given the abundance of splicing regulatory elements, the regulations of splicing are redundant to a certain extent, leading to some atypical splicing variants. This splicing background could increase the diversity of proteins during evolution. I will simulate the mutations in human transcripts to examine how frequent the splice sites and the *cis*-elements can arise from intronic background, thus explore how this mechanism can increase the genomic complexity.

- ***Systematically identify protein trans-factors that regulate splicing***

The *cis*-elements generally regulate splicing by recruiting protein factors (*trans*-factors) that interact favorably or unfavorably with the core splicing machinery. For example, serine/arginine rich proteins usually bind to ESEs and help the assembly of the spliceosome, whereas the heterogeneous nuclear ribonucleoproteins (hnRNP) often interact with ESS to inhibit the use of adjacent splice sites. More than 100 proteins have been identified in the spliceosome (~50% are RNA binding proteins), and many additional protein factors are needed to mediate AS. While interactions between some *cis*-elements and their *trans*-factors are known, most interactions remain to be established.

The following strategies can be used to identify these interactions (Fig. 2). I will first identify *trans*-factors that bind to new ESS, and further extend this exploration into other *cis*-elements obtained through FAS screening.

A). The *cis*-elements will be expressed as a fusion RNA to the MS2 hairpin, a viral RNA that tightly binds to MS2 coat protein. After optional cross-linking of RNA to proteins, the RNA and its associated protein can be purified by the MBP-MS2 fusion protein affinity chromatography (Fig 2A).

B). A yeast three hybrid system will be developed with RNA fused to the MS2 hairpin as an adaptor and Gal4 DNA binding domain fused to MS2 coat protein as bait. For each family of *cis*-elements, we will screen a cDNA library for interaction proteins (Fig 2B).

C). A list of ~1,000 or less RNA binding protein candidates can be chosen through bioinformatic analyses. I will make a protein chip with these candidates and probe it with different RNA to identify the specific interaction (Fig 2C).

D). I will make a reporter minigene similar to that in FAS screen (Fig.1) except inserting a MS2 hairpin in the exon 2, and make stable cell line of such reporter. All three exons will normally be spliced to produce a non-functional gene. I will express the list of candidate *trans*-factors fusing to MS2 coat protein, and screen for proteins that cause exon 2 skipping (*i.e.* producing GFP) upon binding to MS2 hairpin (Fig. 2D).

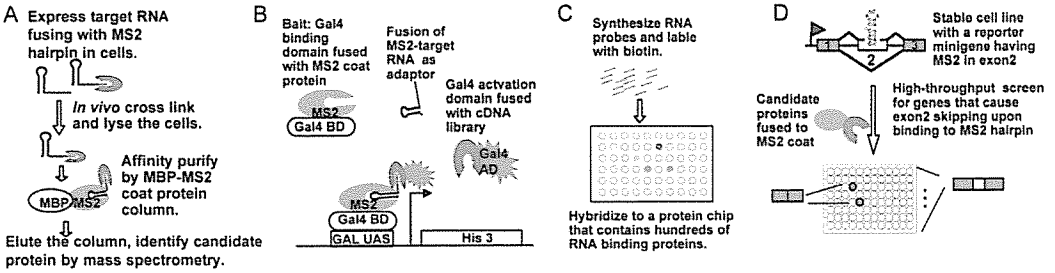

Figure 2. Four proposed strategies to identify the *trans*-factors of splicing.

I will start with strategy A that has been widely used and requires less time. However, since a protein generally binds to its RNA partner with an affinity only several hundred-fold higher than non-specific binding, the low expression proteins could be missed by this strategy. Strategies B and C can normalize for protein levels, thus will solve this difficulty. The protocols for the yeast three hybrid screen are well developed (Bernstein et al, 2002 Methods 26: 123-41) and the cDNA library and yeast strains are commercially available, so it has high possibility for success given adequate time. Developing a high-throughput protein array (Fig. 2C) could be more time-efficient, and the making of such arrays could be simplified with commercially available custom protein array services (like Human ProtoArray of Invitrogen). Strategy D represents a novel functional screen for *trans*-factors recognizing ESS, and can be modified to screen for factors that recognize other *cis*-elements. The combination of multiple strategies makes it possible to identify splicing *trans*-factors in a full scale.

2. Integrate information of *cis*-elements and *trans*-factors to understand splicing regulation

With the knowledge of all splicing regulatory *cis*-elements and their protein factors, I will study how splicing decision is made inside a cell by modeling the interactions between them. Such modeling will eventually make it possible to predict splicing of any transcripts in different tissues.

The first challenge for such modeling is that not all *cis*-elements recruit protein factors inside a cell, i.e., the interactions are context-dependent. This is analogous to that of transcription regulation where only a small portion of transcription factor binding sites identified by sequence searching program are functional. To address this challenge, two approaches will be used to refine the ***in vivo* rules** of how protein factors bind to *cis*-elements: (i) I will identify the biological target of each splicing factor using RNA Immuno-Precipitation combined with microarrays (RIP-Chip). (ii) I will also use RNAi to silence each *trans*-factor, and examine global changes of splicing with alternative splicing microarrays (RNAi-ASChip). Both methods will provide useful information of how *trans*-factors bind to *cis*-elements *in vivo*.

I will collaborate with Burge lab to extract information from the large-scale microarray experiments and to model the interactions between *cis*-elements and *trans*-factors. With such knowledge, I will simulate the splicing decision of any transcripts through following approaches:

A) Forward approach: The relative expression levels of *trans*-factors in each tissue (or cell type) can be obtained through microarray data. These levels will be calibrated for the amounts of proteins with western blot using the available antibodies. Combining such information with the *cis*-elements in each transcript, I will determine the splicing behavior of any transcripts with splicing simulation algorithms like ExonScan (Wang *et. al.* 2004 Cell 119: 831-45).

B) Reverse approach: Since a typical transcript contains multiple splicing *cis*-elements, the alternative splicing of certain genes can reflect the amount (and activity) of splicing *trans*-factors inside a cell. There should be a minimal independent set of genes (or a 'key' set) whose splicing behaviors can represent the splicing environment in a cell. Then the splicing decision of any transcript

can be viewed as a function of splicing behaviors of such 'key' set. I will use computational methods (e.g. machine learning method) to extract the 'key' gene set and solve splicing decision from such set. The RIP-Chip and RNAi-ASChip experiments outlined above can help me to extract such 'key' set.

I expect the experiments and the computational modeling outlined above will provide new knowledge on how splicing is regulated in a systematic level.

3. Screen for chemicals that modulate splicing and could be used as drugs

Given the fact that most human genes contain introns and that the majority of genes are alternatively spliced, it is not surprising that disruption of normal splicing pattern can lead to human diseases. Among the point mutations that cause human genetic diseases, at least 15% of them disrupt splicing (Krawczak et al., 1992 Hum. Genet. 90:41-54). A major class of splicing mutations is the mutation in introns that turns part of an intron into an exon, thus disrupting the protein function. For example, a new exon in intron 11 of CFTR gene is created by a point mutation called 1181+1.6kb A->G, which is the fourth-most-frequent mutations among cystic fibrosis patients in south Spain and France.

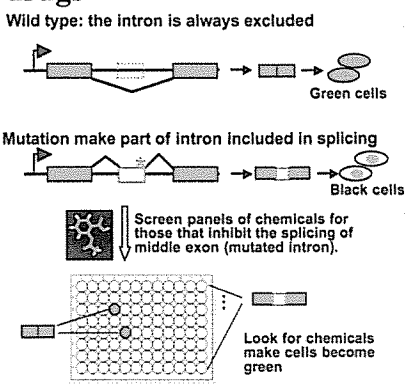

Figure 3. Chemical genetics screen for compounds that can correct splicing mutation.

It has been shown that certain drugs can inhibit splicing of specific exons (e.g., Nissim-Rafinia et al, 2004 EMBO Rep 5: 1071-7). The cell based splicing reporter system I developed provides a platform for a chemical genetic screen of compounds that could correct such splicing mutations (Fig. 3). We will insert the mutated CFTR intron into the GFP reporter system, and generate a stable cell line with this reporter. A library of chemical compounds will then be applied to the cell line using a standard chemical genetic screen protocol. I will then screen for green cells which have recovered normal splicing. The splicing of two GFP exons could serve as the internal control for the specificity of splicing inhibition. As a cell based positive screen, this strategy will also exclude the compounds that lead to cell death. The compounds obtained from the first round of screen can also be subjected to further analysis (or screen) in another splicing construct for their specificity. The lead compounds can also be modified for more specific splicing inhibition.

Initiatives of chemical genetics (such as NIH roadmap initiatives) are being established to collect large library of chemicals and to provide services to scientific community. I will conduct this screen in collaboration with such initiatives. This approach provides a general screen for compounds targeting to a major class of splicing mutations, and is fairly flexible to be adopted for other situations (e.g. screen for compounds to correct intron retention mutations). In addition to selecting lead compounds for drug discovery, this strategy could also produce chemicals that interfere with the splicing machinery, thus providing a tool to study the mechanism of alternative splicing.

Perspective

The major goal of this proposal is to (a) understand features that regulate splicing in a systematic fashion (*i.e.* to clarify the "splicing code") and (b) identify chemical compounds that can correct splicing mutations. The approaches I have outlined form the beginning of a long-term investigation of RNA processing. I believe the experimental and computational methods developed for this purpose could also be used to study other regulatory RNAs. I also look forward to mutually fruitful contacts with colleagues working on fields of both computational and experimental biology. From my experience in Burge lab, it is clear that great science arises from the free flow of ideas between groups in both camps.

Teaching Statement

Zefeng Wang

In my senior year of graduate study at John Hopkins, I had the opportunity to tutor junior graduate students who had problems with courses of molecular biology and cell biology. Since most of these students had undergraduate background other than biology, it is quite a challenge because I have to teach complex biological phenomena or advanced research approaches based on very simple principles. However, this is also intellectually stimulating because I need to think biological question in a different way.

At MIT, I have gained more teaching experience as one of the instructors for undergraduate seminar course (course 7.341, see website listed at the end). This course is for advanced undergraduate focuses on the primary research literature, and the goal is to introduce students with method of contemporary biological research and the logic of experimental design and interpretation. My teaching experiences and research background make me suitable to teach both undergraduate and graduate classes, the following subjects are the examples that I am interested in teaching:

Undergraduate level:

- *Molecular biology, genetics or cell biology.* These courses will teach students the basic concept and logic of contemporary biology, for example, how the genes are expressed and regulated, and how the cellular phenotypes are controlled by genes. The lab works could also be designed on this topic to teach students the techniques of gene expression by transfecting foreign DNA, or gene silence by RNAi.
- *Bioinformatics and genomics.* This course will be targeted to advanced undergraduate. It will cover topics of genomic DNA sequencing, DNA sequence analysis, protein structure analysis and prediction as well as analysis of microarray data. The aim would be get students familiar with the available bioinformatics tools and the current research of computational biology.

Graduate level:

The course I mentioned above can also be taught in graduate level, with more details and more emphases on the design of research project. In addition, I can also teach topics such as:

- *High throughput discovery in molecular biology.* This course will cover the new methods of large-scale discovery, such as RNAi screen, DNA microarray and protein microarry, chemical genetic screen, etc. The students will learn how to design such experiments and how to mine the massive data obtained.
- *Gene regulation by RNA.* This course will cover the role of RNA in the process of gene expression. The primary focus will be the gene regulation in RNA splicing, RNA surveillance and mRNA translation. Other exciting findings in this process, like ribo-switches and microRNA, will also be included.
- *System biology.* This could be a lecture series based on research papers on this evolving field. I will emphasize on the integration of experimental and computational approaches that could elucidate the complex network of gene expression.

Course web pages:

Syllabus and material of our course (RNA splicing and human diseases: molecular and computational approaches) can be found in http://web.mit.edu/people/holste/7.341.html

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Bldg. 68-217, Cambridge, MA 02139-4307

Contact Information for References

Dr. Chris Burge
Department of Biology
Massachusetts Institute of Technology
31 Ames St., 68-223
Cambridge, MA 02139-4307

Telephone: (617) 258-5997
Fax: (617) 452-2936
E-mail: cburge@mit.edu
Web: http://genes.mit.edu/burgelab

Dr. Paul T. Englund
Department of Biological Chemistry
Johns Hopkins Medical School
725 N. Wolfe St.
Baltimore, MD 21205

Telephone: (410) 955-3790
Fax: (410) 955-7810
Email: penglund@jhmi.edu
Web: http://biolchem.bs.jhmi.edu/Englund/index.htm

Dr. Phillip A. Sharp, Ph.D.
Center for Cancer Research
Massachusetts Institute of Technology
40 Ames St., E17-529B
Cambridge, MA 02139-4307

Telephone: 617-253-6421
Fax: 617-253-3867
Email: sharppa@mit.edu
Web: http://web.mit.edu/sharplab/