wellcome trust

Private and Confidential

Dr David Karlin
The Wellcome Trust
215 Euston Road
LONDON
NW1 2BE

Tel: +44 (0)20 7611 8888 Direct: 8284
Fax: +44 (0)20 7611 8545
E-mail: l.barton@wellcome.ac.uk

Our Ref: 090005

15 December 2009

Dear Dr Karlin

Re: **"Discovering the function, structure, and evolutionary impact of proteins created de novo (i.e. not by duplication), in particular in viruses and in bacteria."**

Thank you for attending for interview on 9 December 2009. I am very pleased to inform you that you have been awarded a Career Re-entry Fellowship and would like to welcome you to the community of Wellcome Trust Fellows.

I attach some comments from the referees who reviewed your application, which I hope you will find useful.

An award letter, accompanied by information about how to activate and manage your grant will be sent to you and your Sponsor in due course. At this time, we will also send you further information about your Fellowship.

Congratulations on becoming a Wellcome Trust Fellow. The Trust is keen to foster a close relationship with its Fellows, so please do not hesitate to contact me if you have any queries or if you simply wish to update us on your progress.

Yours sincerely

Louise Barton.

Dr Louise Barton
Grants Adviser
Molecules, Genes and Cells
Grants Management Department

Att: Referees' comments

Cc: Professor Paul Harvey
Professor David Stuart

Referee 1

Candidate

The candidate has a solid and proper training that will allow him to continue successfully with his work. Moreover, he has built a good career and shows a vast interest in the field. The candidate does not have many publications but plays a central role in the ones he has—as a first or last author.

Research Environment

The institution and departments of the candidate's supervisors, Dr. Belashaw and Dr. Grimes, are first class. The experience and impact of this team in the field is outstanding. Therefore, I have no doubts about the quality of the research environment.

Research question

This is an interesting research project that analyzes the sequential, structural, and functional properties of proteins that have been created *de novo* (not by duplication) in virus and in bacteria. The authors propose that the analysis of these proteins will shed light on the role of the *de novo* creation on the cross-species transmission and pathogenicity of virus and bacteria, and on the origin and evolution of protein structural folds in nature. In my opinion, the central research question of this project is important and rather ambitious. Furthermore, the project could help to understand, at least in part, the characteristics of new proteins. It could also yield insights into the process of the origin of proteins.

Research feasibility

The central argument of the project is that the analysis of the structural, and functional properties of *de novo* proteins (not by duplication) in virus and in bacteria, will allow the authors to address two evolutionary issues: a) the relatively small number of protein structural folds in nature and b) the role of de novo protein creation on the cross-species transmission and pathogenicity of viruses and of bacteria. It must be said that the first issue, mentioned above, is a question of enormous proportions. It constitutes a research program that can span many projects. Nevertheless, this work may yield, along the way, important knowledge on this very relevant issue. This can be particularly the case in the second issue we mentioned. Here, the analysis of these proteins promises important data on their role in transmission and pathogenicity. Therefore, the project goals seem feasible even if we recognize that they are quite ambitious.

The bioinformatics milestones are adequate given the timetable. However, some experimental goals are broad for two or three years (i.e. expression, purification and crystallization of 10 *de novo* proteins). Furthermore, milestone number 5 is confusing, it says: "Bioinformatics analysis of overlapping proteins: annotated function". Since the sequence is new, there are no sequences in the bioinformatics databases. Finally, the methodology to identify the overlapping genes and their proteins is not obvious, but previous publications (by the candidate and supervisor) support this point.

Referee 2

Candidate

The candidate showed strong potential in his PhD publishing several good papers from it, relating to disordered viral proteins in 2002-2003. Since then he has worked in other fields promoting science. However, during that time he has managed to do some good preliminary studies relating to the proposed studies with a publication just available (July 2009) in J Virol a top Virology journal.

Research Environment
The environment with the combined expertise of the two groups will provide an ideal environment for the studies.

Research question
The question is certainly original, the idea that new classes of proteins of novel function might be able to be discovered in this way is an important but difficult to test idea. As this proposal aims to study and characterise many proteins its potential to find something very interesting particularly from viruses is large.

Research feasibility
The candidate has a lot of excellent preliminary data. From this there are excellent candidates for proteins having some function. Many of these functions are likely to be novel and it is going to be quite hard to predict a function from the sequences. In addition as the candidate has pointed out these may be unstructured or membrane associated. However the candidate in his PhD project and his supervisors have successfully addressed such technical difficulties before. He has previously shown that such viral proteins are involved in pathogenicity.

Overall Assessment
The candidate's track record and publication history before the break was quite short, but he had three first authorships in two years. He has since held positions of responsibility/leadership in science communication. During this time he has also shown the several hallmarks of successful independent research - in creating the ideas for this project and finding an excellent environment in which the project may be successful. It is a risky project in that when novel proteins are discovered it may be difficult to assign meaningful function or determine structure. However, on the other hand, the discovery of even a few new classes of (viral) proteins would be a substantial advance.

Referee 3

Candidate
Following a highly successful PhD (including 3 first author publications), David left research for science communication, though continued to collaborate and publish (including 2 senior author publications) as a free-lance scientist. After experiencing another side of science, David now wishes to return to full-time research. David's career to-date exhibits a considerable degree of independence and drive, besides scientific excellence. He appears to be an ideal candidate for a Wellcome Trust Career Re-Entry Fellowship.

The proposed research will allow David to build upon, and extend, his current research interests. One particularly attractive aspect of the proposal is the combination of bioinformatic analysis and experimental follow-up. Computational analysis is an integral part of modern biological science, and researchers who are adept at both computational and experimental approaches have a distinct advantage. David already has considerable experience in both areas, but the proposed research will give David the opportunity to further extend and enhance his skills in statistics and bioinformatics. David's proposed host institution is second to none in the fields relevant to David's proposed research - thus providing David with daily access to leading researchers and an ideal location for his career development. The candidate's plan to organize and host an international workshop on de novo protein creation shows initiative and, given the candidate's previous experience, it is likely to be both successful and to further establish the candidate's already strong reputation in the field.

Research Environment
The University of Oxford is one of the best locations in the world for the proposed research. It has a very strong record in evolutionary and mathematical biology including - of particular relevance to the proposal - both virus bioinformatics and analysis of overlapping genes, with a significant number of researchers in fields of direct relevance to David's proposal. David's proposed bioinformatics supervisor, Robert Belshaw, is a leading researcher in virus bioinformatics and has recently published a seminal article on the evolution of overlapping genes in RNA viruses. A number of other researchers in relevant fields are easily accessible elsewhere in the UK. I can think of no better place for David to pursue his research. The University of Oxford also offers an ideal environment - including a variety of courses - for David to continue developing his career as an independent scientist and future lab head.

Research question
This is an innovative and original proposal that addresses a number of important, but little-studied, questions, for example 'What are the characteristics of overlapping genes from a protein point-of-view?' (previous studies on the global characteristics of overlapping genes have been mainly restricted to the nucleotide rather than amino acid sequence), and 'What proportion of de novo genes arise from out-of-frame overprinting of preexisting genes and what are the consequences of this for the evolution of new protein folds?', besides improving our understanding of the nature and functions of proteins (both disordered and ordered) encoded by overlapping genes. As such, there is no doubt that the proposed research will lead to a good number of novel and exciting discoveries and publications, and provide significant and important advances in a number of fields.

The candidate is well aware of the dangers of over-extrapolating the results on virus overlapping genes (where evolution of the de novo protein is restricted by the requirement for coding in the overlapping ancestral gene) to cellular de novo proteins (where after initial 'creation' via 'overprinting' of an existing gene, duplication of an overlapping pair may subsequently allow the ancestral and de novo proteins to evolve independently). Even so, I believe that the research will provide real insights into the evolution of cellular de novo genes - which is an important component of understanding the origin and evolution of proteins per se. The proposed extension of the analysis to include a sample of bacterial overlapping genes is an integral part of this.

Aside from providing new insights into gene evolution and the characteristics of de novo and/or overlapping proteins, the research will also provide new structural and functional information for a selection of currently poorly characterized viral accessory proteins. This alone is a valuable research outcome, and should result in several publications.

The union of David Karlin's and bioinformatic supervisor Robert Belshaw's complementary ideas and experience, tied together with common interests in virus overlapping genes and evolutionary processes, should prove to be a powerful and synergistic combination in terms of results and new insights.

Research feasibility
There is no doubt in my mind that the bioinformatics component of this proposal is feasible and attainable within the time-frame set out. David has recently published an article that essentially serves as a 'pilot study' for the bioinformatics component of the proposed. There should be no problem extending the methodology therein to the full set of known overlapping genes in viruses (and some from bacteria), and the extension will no doubt yield many new and interesting results, besides an appropriate number of candidates for the experimental follow-up component of the proposal. The bioinformatics analysis will be done in conjunction with David's proposed supervisor, Robert Belshaw, whose recent publications (e.g. 'The RNA virus database' and 'The evolution of genome compression ... in RNA viruses') clearly indicate that he has the necessary expertise to provide David with any further support

needed for this analysis.The timescale, milestones and expected numbers of candidates etc appear realistic and achievable. The methodology and design are well thought out and clearly described. The requested resources (person-years and computer equipment) appear reasonable.

The following are a few minor comments that the candidate may wish to take on board but are in no way intended to detract from the proposal as it stands:
1) I note that the candidate has (in my opinion, wisely) decided not to include bacteriophages in the current proposal. He might also wish to exclude other large dsDNA viruses (I suspect the effort to payoff ratio may be relatively low for these viruses), though certainly small dsDNA viruses are worth including.
2) I suspect that the candidate is likely to find bacterial genomes a reasonably rich source of new overlapping genes, including a fair number of long overlaps (as opposed to short terminal overlaps), many of which are currently unannotated but fairly accessible to bioinformatic prediction with comparative methods. The candidate only requires 50 bacterial overlaps and this should be achievable with just published examples, as outlined in the proposal. Further analysis is probably beyond the scope of the current proposal but could make a nice future or additional project - and may even provide some cases where a de novo gene overlaps an ancestral gene in one species but, in another species, gene duplication has allowed the de novo and ancestral genes to evolve independently.

Overall Assessment
The proposal builds on and extends the candidate's previous work and should provide diverse and new exciting results relevant to protein evolution, protein structure, virology and virus bioinformatics. Both the host institution and the proposal provide excellent career development opportunities. In particular the candidate's plan to develop and combine both bioinformatic and experimental lines of research will place him in an excellent position to perform world-class research and, at the end of the fellowship, establish his own lab.

Referee 4

Candidate
The candidate is very suitable for a career re-entry fellowship. He was a very productive researcher during his PhD period. Moreover, he wrote two papers on his own time while not being a professional scientist. The latter demonstrates that he is serious about his desire to go back to science, and that he can work independently without having a formal supervisor. I believe that this fellowship will allow the applicant to jump-start a successful career in science and to carve a niche for himself.

Research environment
The research environment is superb. It would have been difficult for the candidate to choose a better place in the UK to carry out the proposed research.

Research question
The research question is highly original and has high potential impact. We still have a very poor understanding of where new protein structures come from. Looking at overlapping reading frames is a great way to learn about de novo creation of proteins, because one of the two proteins must have evolved de novo. To my knowledge, very little work has been done on this specific question.

If everything goes according to plan, the research could be very high impact. In the best-case scenario, if the applicant can demonstrate that the majority of the newly solved structures have little similarity to existing structures, that would show that there is apparently a lot of room for new protein structures to arise. Alternatively, if many of the structures show

significant similarity with already existing structures, that would also be interesting. In this case, the conclusion would be that the space of viable protein structures is rather limited and evolution rediscovers the same structures over and over again. Both of these outcomes would be major advances to our understanding of the origin of protein structures.

Research feasibility
Overall, I think the proposed research is feasible. The requested resources are appropriate for the expected outcome of the project.

Strengths:
The research project is fairly diverse, with several different approaches (both computational and experimental) that can be carried out independently. If one subproject doesn't quite work out as planned, none of the other subprojects are affected.

Weaknesses:
The goal of determining the structure of 10 proteins is ambitious. The applicant may not be able to complete this goal. I consider this a minor weakness, because the project will produce worthwhile insight even if only a smaller number of structures can be obtained. The computational part of the project is independent of this weakness and can be carried out in any case.

Overall Assessment
The proposed research is innovative and exciting. The applicant is well qualified to carry out the work and has chosen a superb environment to complete the project.

Referee 5

General Comments:
This project is aimed at understanding the evolvability of proteins and the generation of de-novo proteins in viruses. The potential benefit of the project is twofold: understanding the fundamental processes underlying the emergence of biological complexity; and stating the fundamental basis for the development of virus proteins recognition. One of the aspects the applicant has not dealt with is how this project in fact could indeed help learning the basis for protein engineering. This would be of unprecedented potential in biotechnology and biomedicine.
This project has several difficult sections such as protein crystallization and identification of potential de-novo protein formation in viruses. Because of this and other question I'll provide some possible aiding points to try to clarify this project. I would like to stress that these points are in no way criticism to the project. This project therefore deals with an interesting but hitherto obscure and very little understood subject.

1. **Comments to the project:**
    a) Protein folding is an important problem in biology and our inability to precisely predict protein folds is testament to our sketchy knowledge about the subject. It is in this point where the candidate would make a fundamental contribution in developing a more realistic approach to protein folds determination. As I said before, the link between this and pathogenesis (viral pathogenesis) would provide the project by a biomedical dimension as well as theoretical. The candidate should therefore consider this as an achievable option or at least as an option to consider despite performing an ambitious objective.
    b) Candidate states that he will be solving 10 three-dimensional structures. I presume that he already has these 10 candidates because otherwise the objective can be optimistic. Solving 3D structure has always been a problem with unpredictable solutions, in special when dealing with unstructured proteins

c) For the sake of formality, please replace the word “creation” by “emergence”. The project is aimed at understanding the **evolution** of these proteins and therefore I assume that there is no intention of finding the creator.
d) There is some work showing that proteins evolve towards the threshold of biologically”energetically” acceptable stability. This makes proteins having an important potential evolvability. I would encourage the candidate and collaborators to perform some stability experiments (WESTERN Blotts, etc.) to see if the proteins emerging de-novo have also evolved in this way. My prediction is that this is the case and in fact this may provide pathogens to evolve quickly to colonize highly specialized niches.
e) I would suggest establish (at least state more clearly in the grant proposal) a good collaboration with bioinformaticians to learn about the problems of determining proteins structures in unstructured proteins and aid developing new computational tools for such purpose. Candidate assures he will develop a database where the information will be publically available. I would ask the candidate to give more information as to how he is intending to develop the database and what kind of information will this contain.
f) I would also ask the candidate explain how feasible his future collaborations are. Hopefully, the objectives of the project will not depend entirely upon these collaborations.
g) Finally, what would the candidate do if some of the objectives fail? For example, does the candidate have alternative approaches, solutions, etc?

2. Comments to the candidate

I find the candidate to have a very interesting profile. Certainly he has a very particular academic profile and I think it is incredibly appropriate to achieve the different objectives lied out throughout the project. The candidate has a solid formation in molecular biology and clear vocation for science. I would hence support the adequacy of the candidate to perform this project despite the gaps in his publication record. Not only the candidate seem to be a good researcher but also a good communicator and hence I would say he would be perfect ambassador from Welcome Trust to the public to diffuse science and knowledge. The candidate also has important collaborations and the mentors and supervisors clearly support him and are very eager to go ahead with the objectives of the project.

3. Comments to deliverables and budget

Regarding deliverables, I find the objectives realistic based on his experience and on the collaborations. I would however give more details regarding the molecular experiments and deliverables in each of the years.

In sum, I find the project exciting, the candidate appropriate and the study and performance of the objectives well thought. The study is interesting from the theoretical as well as biomedical points.