CAREER 05/07
eGrants – Version 1

APPLICATION FOR A CAREER RE-ENTRY FELLOWSHIP (JUNIOR)

Trust reference number: 046937
eGrants reference number: 12273

Q1	Applicant
Surname	Karlin
Forenames	David
Title (Dr etc.)	Dr

	Sponsor	Sponsor	Sponsor
Surname	harvey	Stuart	
Forenames	paul	David Ian	
Title (Dr etc.)	Prof	Prof	

	Supervisor	Supervisor	Supervisor
Surname	Belshaw	Grimes	
Forenames	Robert	Jonathan	
Title (Dr etc.)	Dr	Dr	

Q2 Title of project:

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

Q3 Department name and address of administering institution where different from applicant's address:

Dpt of Zoology
The Tinbergen Building
University of Oxford
South Parks Road
Oxford
OX1 3PS
United Kingdom

Q4 Period for which support is sought: (state in months) 36

Q5 Proposed start date: 01/01/2010

Applicant

Name: Dr David Karlin

Contact address:
MSH

The Wellcome Trust

London
NW1 2BE
United Kingdom

Telephone numbers:

Day: 020 7611 7350

Mobile:

Fax.:

email: karlin.david@gmail.com

Supervisor

Name: Dr Robert Belshaw

Contact address:
Zoology

University of Oxford

Oxford
OX13PS
U.K.

Telephone numbers:

Day: 018 6528 1997

Mobile:

Fax.:

email: robert.belshaw@zoo.ox.ac.uk

Supervisor

Name: Dr Jonathan Grimes

Contact address:
Nuffield Dept of Clinical Medicine

University of Oxford

Oxford
OX37BN
UK

Telephone numbers:

Day: 018 6528 7561

Mobile:

Fax.:

email: jonathan@strubi.ox.ac.uk

Sponsor

Name: Prof paul harvey

Contact address:
zoology

University of Oxford
south parks road
oxford
OX1 3PS
england

Telephone numbers:

Day: 018 6527 1260

Mobile:

Fax.:

email: paul.harvey@zoo.ox.ac.uk

Sponsor

Name: Prof David Ian Stuart

Contact address:
Division of Structural Biology

University of Oxford

Oxford
OX3 7BN
UK

Telephone numbers:

Day	004418 6528 7567
Mobile	4477 4777 8107
Fax.	004418 6528 7547
email	Dave@strubi.ox.ac.uk

Q6 CURRICULUM VITAE OF APPLICANT

(a) Surname: Karlin Forenames: David

Date of birth: 17/03/1973 Sex: Male Nationality: French

(b) Title of current post (If unemployed or in temporary employment, please give details of last appropriate post):

Project Manager, Public engagement

Date of appointment/Start date of last appropriate post: 13/01/2009
01/07/2007

Expected date of termination/End date of last appropriate post: 31/12/2009
14/09/2008

(c) With whom do you have your contract of employment?

The Wellcome Trust

(d) Current/last appropriate salary details

Salary grade:	5
Basic salary:	39000
London Allowance:	
Salary enhancements:	
Currency:	Other
Please specify currency (If 'Other'):	euro
Date of last increment:	01/07/2007

Source of personal salary support (If 'Other', please specify):

Please also be specific if salary is funded from more than one source.

The Wellcome Trust

(e) Previous posts held: (most recent first)

Date from	Date to	Position	Department	University/ Institution
2007	2008	Director	Public Programmes	Tous Chercheurs (popular science association), Marseilles, France

2004	2008	Freelance bioinfo	-	No institutional affiliation. Marseilles, France
2002	2006	Director	-	DNA School (popular science association), Marseilles, France
1997	1998	Molecular virologist	Virology Dpt	Institute of Tropical Medicine, (popular science association), Marseilles, France

(f) Education/training:

Date (mm/yyyy)	Degree	Subject	University/Institution
2002	PhD	Structural Virology	University of Marseilles, France
1997	MSc	Cellular and Molecular Pharmacology	Paris VI University, France
1996	Engineer Degree	Process Engineering	Ecole des Mines de St-Etienne (France) and University College Dublin (Ireland)
1993	BSc	Maths-Physics-Chemistry	Lycee Henri IV, Paris, France

(g) Higher degree

(i) Are you registered for a higher degree, e.g. PhD, MD or equivalent? No

(ii) If yes, please specify degree, university and likely date of completion.

(iii) If no, do you intend to register for a higher degree if awarded a fellowship? No

(h) Summary of scientific career to date, including key achievements (no more than 700 words).

After a MSc in molecular and cell biology, I did my military service in the institute of tropical medicine of Marseilles, which prompted a life long interest in viruses. I stayed in Marseilles for my PhD in which I was supposed to tackle a new research project: solving the 3D structure of proteins of the replicative machinery of measles virus. However, all the proteins I worked on were either proteolysed, aggregated, or not crystallisable, or all of this at the same time. I discovered that the common reason for this behaviour was that they were all unstructured. This initiated a new strand of research in the lab, focusing on unstructured proteins. During the course of this research I gained a **considerable amount of experience in protein expression, purification and characterization**.

I loved science but thought there was more about it than working in a lab and publishing papers, and wanted to see how science was developing as a societal activity. In particular, I had always harboured a strong interest in public engagement and after my PhD in 2002, I set up a popular science association hosted within a research institute. It included a molecular biology teaching lab and I **kept my bench skills up to date** by teaching vocational training in molecular biology and biochemistry.

These years spent doing public engagement allowed me to take a step back and to acquire a good background in human genetics and in many areas of biology. They also gave me a broader perspective and understanding of research and of its context, including its applications in health. For instance I created workshops on biomedical research for patient groups, which created, or strengthened links between disease associations and researchers/clinicians. I met numerous inspiring researchers and learned an awful lot on how good science could be done.

This deeper understanding revived my interest for an intriguing observation made at the end of my PhD: almost all unstructured proteins I had identified in measles virus were encoded by overlapping genes. Thus, I enlisted the help of a bioinformatician from my former lab and of American specialists of the prediction of unstructured proteins. In my spare time (the advantage of bioinformatics research is that only a computer is needed), I carried out and coordinated a work that confirmed the initial intuition and evolved into the more general question of the creation of novel proteins by viruses. I authored two papers, one on disorder prediction and the other one on the research project itself (in minor revision at J Virol, May 2009). This research thus constitutes the equivalent of two year's postdoctoral experience.

The submission of the article coincided with my moving to the UK for familial reasons. There I met Robert Belshaw and Jonathan Grimes in Oxford who had expressed ideas complementary to mine. We shared a clear sense of a unique opportunity to make a significant contribution to the emerging field of de novo protein creation, both from an evolutionary and from a structural point of view. Working now at the Wellcome Trust, I can see first-hand the excellence of the UK research and I wish to return to my main professional interest, research, in the UK, as my career.

I feel I could have a significant impact on my field, thanks to the skills learned in my previous experience (coordination with different teams, organisation of a workshop to highlight this novel field of research, grant writing and excellent time management skills) and thanks to the broader scientific outlook I have gained. For these reasons and because the project is very innovative, I think a career re-entry grant from the Wellcome Trust would be ideally suited.

(i) If a fellowship is awarded, in what way will this further the applicant's career? (no more than 300 words)

Working now in science communication in the UK, I can see first-hand the excellence of the UK research and I wish to return to my main professional interest, namely research, here in the UK.

This multidisciplinary project is well suited to my abilities, both scientific (a combination of bioinformatics, biochemistry and virology), and managerial (strong time and project management, grant writing and communication), the latter acquired during my career break from research.
Working in excellent laboratories with different but complementary interests will greatly further my professional skills. Curating and analysing hundreds of viral genes in depth will give me a precious opportunity to further my knowledge of biological pathways. Thanks to the available training, I will learn the latest concepts and analysis techniques in evolutionary biology, and state-of-the-art techniques in structural biology.

I will organise an international workshop to help coalesce a novel field of research (de novo protein creation), which should open up many opportunities for collaboration with other researchers (see Data management and sharing). For all these reasons, I feel I could have a significant impact on the field, and after the fellowship, go on and develop more of my own research projects.

In summary, a career re-entry grant would offer me great opportunities to pursue a successful research career; reciprocally, the Wellcome Trust can be ensured of my drive and motivation, which I proved by doing unfunded research in my spare time for 6 years, publishing 2 articles in good journals.

(j) What scientific considerations led you to choose this laboratory and supervisor for your research? If you have already been based in this laboratory for a year or more, please specify your reasons for remaining (no more than 400 words).

Both Robert Belshaw and Jonathan Grimes have expressed ideas complementary to mine (see Background). We share a clear sense of a unique opportunity to make a significant contribution to the emerging field of de novo protein creation, both from an evolutionary and structural point of view, by putting our strengths in common.

- R. Belshaw is very experienced in the bioinformatics study of viral evolution and has led the first ever systematic analysis or the evolution of overlapping genes. His quality is attested by the Wellcome Trust project grant he has recently received. He still programs, has employed a PhD who is a bioinformatician and thus his team will provide me with a very favourable bioinformatics environment. His teaching skills, life and research experience make him an ideal supervisor.

- J. Grimes is based in one of Europe's leading structural biology labs, with access to the most up-to-date techniques and facilities. Having been briefly in contact with him during my PhD (our respective labs worked together), I know he is a very creative and skilled structural biologist, still actively involved in the solving of X-ray structures, and passionate about uncovering the function of viral machineries.

(k) Publications

Please list all publications, including original research publications and other scholarly contributions. Publications should be in chronological order with the most recent first. **Please give citation in full, including title of paper and all authors**.

PUBLICATIONS AND COMMUNICATIONS of David Karlin

Peer-reviewed articles

The bioinformatics work corresponding to the two articles below was done and coordinated on my spare time while working in public engagement.

1. Rancurel C., Pedro R., Khosravi M., Canard B., Dunker K., **Karlin D**. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *In revision for Journal of Virology* (May 2009).
   A figure summing up the main findings is available as Additional Document 1

2. Ferron F, Canard B, Longhi S, **Karlin D**. A practical overview of protein disorder prediction methods. *Proteins*. 2006 Oct 1;65(1):1-14.

3. **Karlin D**, Ferron F, Canard B, Longhi S. Structural disorder and modular organization in Paramyxovirinae N and P. *J Gen Virol*. 2003 Dec;84(Pt 12):3239-52.
4. Longhi S, Receveur-Brechot V, **Karlin D**, Johansson K, Darbon H, Bhella D, Yeo R, Finet S, Canard B. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem.* 2003 May 16;278(20):18638-48.
5. **Karlin D**, Longhi S, Canard B. Substitution of two residues in the measles virus nucleoprotein results in an impaired self-association. *Virology*. 2002 Oct 25;302(2):420-32.
6. **Karlin D**, Longhi S, Receveur V, Canard B. The N-terminal domain of the phosphoprotein of Morbilliviruses belongs to the natively unfolded class of proteins. *Virology*. 2002 May 10;296(2):251-62.

Other publications

7. 10 simple rules for conveying science to the general public. Bourne PE, **Karlin D**. (submitted to Plos Computational Biology).
8. Receveur-Bréchot V, **Karlin D**. Le désordre, clé de voûte des protéines. La Recherche, 2005 June, 387, 52-55

Communications

2008	PFAM team, Sanger Institute	Cambridge (UK)	Talk
	Structure and properties of proteins created de novo in viruses		
2005	French Biophysics Society Congress	Anglet (France)	Talk
	A practical overview of disorder prediction methods		
2002	International Society of Microbiology Congress	Paris (France)	Poster
	Structural disorder and modular organization in Paramyxovirinae N and P		
2000	American Society of Virology Congress	Fort-Collins (USA)	Poster
	The N-terminal domain of the Morbillivirus phosphoprotein is disordered		

Invited stay

Keith Dunker's bioinformatics lab (Indianapolis, USA, 2 weeks in 2004) for a collaboration on the structural properties of proteins encoded by overlapping genes. It led to publication 1) above in which I was last author.

Q7 CLINICAL STATUS

THIS SECTION MUST BE COMPLETED BY ALL APPLICANTS WHO ARE MEDICAL OR DENTAL GRADUATES

(a) What level of clinical contract do you currently hold? If 'Other', please specify.

(b) Name of Health Authority or Hospital Trust:

(c) Date current contract expires:

(d) Please state your chosen clinical specialty, if known:

(e) What progress, if any, has been made towards accreditation in your chosen specialty?

(f) i) Do you hold a National Training Number (NTN)?

ii) If yes, state NTN and date awarded:

iii) If no, when do you intend to apply for a NTN?

iv) In which postgraduate deanery is your NTN held, or will be held?

(g) i) Do you hold a Certificate of Completed Specialist Training (CCST)?

ii) If yes, state date awarded:

iii) If no, what date would you expect to qualify to receive your CCST, assuming your fellowship application is successful? (mm/yy)

(h) What level of honorary clinical contract will be sought during this award? If 'Other', please specify.

(i) i) Please state the clinical duties that are essential for the proposed research and the time required each week to perform these duties:

(i) ii) Please state what clinical duties are essential for the minimum requirements for higher training in your specialty, and how you intend to meet them:

(i) iii) Please state the total time you intend to spend each week on clinical work, including (i) i) and (i) ii) above:

Q8 RESEARCH GRANTS FROM OTHER FUNDING AGENCIES

Research grants held in the last five years and any key prior grants (list the most recent first). Please state the name of the awarding body, title of project, amounts awarded and start and end dates of support. For all current grants, indicate the number of hours per week that are spent on each project.

Key prior research grants:

Oct 2001-May 2002: PhD Extension Grant of the Fondation pour la Recherche Medicale (Medical Research Foundation, Paris, France)

1998-2001: PhD fellowship specific for engineers, CNRS (Centre National pour la Recherche Scientifique, Paris, France)

Q9 PREVIOUS APPLICATIONS TO THE WELLCOME TRUST

(a) Is this the Applicant's first application to the Wellcome Trust? Yes

(b) Give details of all previous applications to the Wellcome Trust over the last five years. Please include name of grant holder, grant number (if known), title of project and, if application was successful, the amount and period of award.

Q10 RELATED APPLICATIONS ***(Please ensure that you read the guidance notes before completing this section, and that you comply with all the stated requirements while your Wellcome Trust application is being considered.)***

(a) Is this or a related application currently being submitted elsewhere? No

If yes, to which organisation?

By what date is a decision expected?

(b) Has this, or a similar, application been submitted elsewhere over the past year? No

If yes, to which organisation?

What was the result?

(c) Is this application a resubmission or has it been previously considered under another Wellcome Trust scheme? No

If yes, when was it originally considered?

Please give the Wellcome Trust's reference number:

State how this application differs from the original (no more than 500 words).

Q12 CURRICULUM VITAE OF SUPERVISOR

(a) Surname: Belshaw Forenames: Robert

(b) Title of current post: Departmental Lecturer

Date of appointment: 26/02/2007

Expected date of termination: 28/02/2014

(c) With whom do you have your contract of employment?

University of Oxford

(d) Source of personal salary support (If ‘Other’, please specify):

Please also be specific is salary is funded from more than one source.

University of Oxford

(e) Previous posts held: (list the most recent first)

Date from	Date to	Position	Department	University/Institution
02/2007	present	Department Lecturer	Zoology	University of Oxford
08/2005	01/2007	Post-Doctoral Research Associate	Zoology	University of Oxford
01/1994	07/2005	Post-Doctoral Research Associate	Biology	Imperial College London
01/1988	12/1993	Research Fellow	Entomology	The Natural History Museum, London

(f) Relationship of current application to other work in the supervisor’s laboratory (no more than 500 words)

I became a lecturer at Oxford in 2007 and, following the success of my first grant application, my first post-doctoral research assistant has just started working with me. This Wellcome Trust-funded project investigates the proliferation of endogenous retroviruses by analysing all (over 400) vertebrate published genome sequences (or sequence fragments), continuing a line of bioinformatic research that I developed prior to moving to Oxford.

Since moving to Oxford I have extended my interest into the broader field of genome evolution. In my opinion piece in Trends Ecol. Evol. last year I outlined my views on the major evolutionary forces determining the architecture of virus genomes. One of these is genome compression, one aspect of which is gene overlap. I have recently developed an evolutionary model for this process (published in Genome Res.) and am about to submit a further paper in which I compare the amount of gene overlap across all known viruses. This will allow me to test the varied hypotheses that have been proposed to explain why genome compression is some common in viruses.

My work with David will be a new way of looking at gene overlaps: investigating the characteristics of novel proteins produced by this process. As David discusses in his application, gene overlaps are a major source of novel proteins, one that has received very little research compared to the mechanism of gene duplication. The collaboration of Johnathan Grimes in David's project is an exciting merger of bioinformatic and experimental lines of research; I am confident that it will reveal new insights into how proteins evolve and am very enthusiastic to be involved.

David's previous research in this field involves mainly very detailed analysis of a relatively small collection of instances of gene overlap in viruses, whereas my work is entirely automated and focuses on the databasing and statistical analysis of very large datasets. We will combine the best of our approaches to generate a large curated dataset and work together on manipulating and analysing it. My new postdoc on the Wellcome Trust grant (Aris Katzourakis) is also an experienced bioinformatician, thus providing further support for David here in the department.

Finally, an additional line of my research in which David's work and expert advice will be very useful is the production of automated tools for detecting and analysing overlaps (Recent work by Andrew Firth has shown that they are even more common in viruses that previously shown, and that some overlaps which play a crucial functional role for the virus have been overlooked). David's library of curated gene overlaps will serve as a benchmark for these software. For maximum usefulness for the community, his library will be linked in the website of my database called the RNA Virus Database (http://virus.zoo.ox.ac.uk/rnavirusdb), just published in Nucleic Acids Res., which I intend to develop as a hub for more specialised virus websites.

(g) Recent publications
No more than **ten** publications which you consider the most important and relevant to this application. Publications should be in chronological order with the most recent first. **Please give citation in full, including title of paper and all authors**.

1. Belshaw, R., de Oliviera, T., Markowitz, S. & Rambaut, A. (2009). The RNA Virus Database. Nucleic Acids Research. 37: D431-D435.

2. Belshaw, R., Gardner, A., Rambaut, A. & Pybus, O.G. (2008). Pacing a small cage: mutation and RNA viruses. Trends in Ecology and Evolution. 23: 188-193

3. Belshaw, R., Pybus, O.G., Rambaut, A. (2007). The evolution of genome compression and genomic novelty in RNA viruses. Genome Research 17: 1496-1504.

4. Belshaw, R., Watson, J., Katzourakis, A., Howe, A., Woolven-Allen, J., Burt, A. & Tristem, M. (2007) Rate of recombinational deletion among human endogenous retroviruses. Journal of Virology 81: 9437-9442.

5. Belshaw, R. & Bensasson, D. (2006) The rise and falls of introns. Heredity 96: 208-213.

6. Belshaw, R., Dawson, A., Woolven-Allen, J., Redding, J., Burt, A. & Tristem, M. (2005) Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present day activity. Journal of Virology 79: 12507-12514.

7. Belshaw, R., Katzourakis, A., Paces, J., Burt, A. & Tristem M. (2005) High copy number in human endogenous retrovirus (HERV) families is associated with copying mechanisms in addition to re-infection. Molecular Biology and Evolution 22: 814-817.

8. Belshaw, R. & Katzourakis, A. (2005) BlastAlign: a program that uses blast to align problematic nucleotide sequences. Bioinformatics. 21: 122-123.

9. Belshaw, R. Pereira, V. Katzourakis, A., Talbot, G., Paces, J., Burt, A. & Tristem M. (2004) Long-term re-infection of the human genome by endogenous retroviruses. Proceedings of the National Academy of Sciences of the United States of America 101: 4894-4899.

(h) Research grants held by supervisor
Research grants held in the last five years and any key prior grants (list the most recent first.
Please state the name of the awarding body, title of project, amounts awarded and start and end dates of support. For all current grants, indicate the number of hours per week that are spent on each project.

Grants from other funding agencies

Grants from the Trust
Grant holder: Robert Belshaw
Grant number: 086173/A/08/Z
Grant title: The evolution and present day activity of endogenous retroviruses (Project Grant)
Successful: Yes
Amount: £217,285
Period of award: 01/05/2009 – 30/04/2012

Q12 CURRICULUM VITAE OF SUPERVISOR

(a) Surname: Grimes Forenames . Jonathan

(b) Title of current post: Research Lecturer

Date of appointment: 01/10/2004

Expected date of termination: 01/10/2014

(c) With whom do you have your contract of employment?

(d) Source of personal salary support (If ‘Other’, please specify):

Please also be specific is salary is funded from more than one source.

HEFC

(e) Previous posts held: (list the most recent first)

Date from	Date to	Position	Department	University/Institution
1/10/1999	31/9/2007	Royal Soc Research Fellowship	NDM	Oxford University
1/9/1998	31/9/1999	Post-doctoral Research Fellow	Biochemistry	Oxford University
1/10/1996	31/8/1998	EMBO Fellows	Grenoble	EMBL
1/9/1993	31/8/1996	Post-doctoral Research Fellowship	Biochemistry	Oxford University

(f) Relationship of current application to other work in the supervisor’s laboratory (no more than 500 words)

The central theme of my research interests is understanding how viruses function, and in particular how they interact with the cells they infect.
Alongside studies on cell entry and exit, as well as macromolecular assembly, I am interested in how viruses modulate the response of the host cell they infect.

It is becoming increasingly apparent that viral genomes contain genes that encode non-essential proteins (ie not proteins that are key to viral replication) and that these proteins play a crucial role in viral pathogenesis.
David has shown from his preliminary work that many of these viral proteins have arisen recently (evolved de-novo and not by gene duplication) and by mechanisms that are not well understood. Thus our research interests are synergistic and complementary, since he will be able to identify newly evolved viral proteins that have roles in viral pathogenesis and immunomodulation.

I use X-ray crystallography as well as other biophysical techniques to study viral proteins and macromolecular complexes. Using these methodologies I was involved in the structure determination of a newly evolved protein from the SARS virus. The structure was problematic to solve, primarily due to the high degree of disorder in the structure. We hypothesised that this disorder may be a feature of newly evolved proteins. So when David approached me with the proposal to study a number of "newly-evolved" viral proteins I was very keen for him to come to the Division of Structural Biology and become fully engaged!!

(g) Recent publications

No more than **ten** publications which you consider the most important and relevant to this application. Publications should be in chronological order with the most recent first. **Please give citation in full, including title of paper and all authors**.

Bamford, D.H., Grimes, J. M. and Stuart, D. I. What does structure tell us about virus evolution? Curr Opin Struct Biol. (2005) 15, 655-63.

Meier, C. M., Aricescu, A. R., Assenberg, R., Aplin, R. T., Gilbert, R. J. C., Grimes, J. M. and Stuart, D. I. The crystal structure of ORF-9b, a lipid-binding protein from the SARS coronavirus. Structure, (2006) 14, 1157-65.

Salgado, P. S., Koivunen, M. R. L., Makeyev, E. V., Bamford, D. H., Stuart, D. I. and Grimes, J. M. An evolutionary link between RNA silencing and transcription revealed by the structure of an RNAi polymerase. Plos Biol. (2006) 4, 2274-2281..

Cooray, S., Bahar, M. W., Abrescia, N. G. A., McVey, C. E., Bartlett, N. W., Chen, R. A.-J., Stuart, D. I., Grimes, J. M. and Smith, G.L. Functional and structural studies of the vaccinia virus virulence factor N1 reveal a Bcl-2-like anti-apoptotic protein. J. Gen. Virol. (2007) 88, 1656-1666.

Graham, S. C., Bahar, M. W., Abrescia, N. G. A., Smith, G. L., Stuart, D. I., and Grimes, J. M. Structure of CrmE, a virus-encoded tumour necrosis factor receptor. J. Mol. Biol. (2007) 372, 660-671.

Bahar, M.W., Kenyon, J.C., Putz. M.M., Abrescia, N.G.A., Pease, J.E., Wise, E.L., Stuart, D.I., Smith, G.L. and Grimes, J.M. Structure and Function of A41, a Vaccinia Virus Chemokine Binding Protein. Plos Path. (2008) 4, 55-68.

Graham, S.C., Bahar, M.W., Cooray, S., Chen, R.A., Whalen, D.M., Abrescia, N.G., Alderton, D., Owens, R.J., Stuart, D.I., Smith, G.L. and Grimes, J.M. Vaccinia virus proteins A52 and B14 Share a Bcl-2-like fold but have evolved to inhibit NF-kappaB rather than apoptosis. PLoS Path (2008) 4: e1000128 doi:10.1371/ journal.ppat.1000128

Abrescia, N.G.A., Grimes, J.M., Kivela, H.M., Assenberg, R., Sutton, G.C., Butcher, S.J., Bamford, J.K.H, Bamford, D.H. and Stuart, D.I. Insights into Virus Evolution and Membrane Biogenesis from the Structure of the Marine Lipid-Containing Bacteriophage PM2. Mol Cell (2008) 31, 749-761.

Poranen, M.M., Salgado, P.S., Koivunen, M.R., Wright, S., Bamford, D.H., Stuart, D.I. and Grimes, J.M. Structural explanation for the role of Mn2+ in the activity of phi6 RNA-dependent RNA polymerase. Nucleic Acids Res. (2008) 36(20), 6633-6644.

Graham, S.C., Assenberg, R., Delmas, O., Verma, A., Gholami, A., Talbi, C., Owens, R.J., Stuart, D.I., Grimes, J.M. and Bourhy, H. Rhabdovirus Matrix Protein Structures Reveal a Novel Mode of Self-Association. PLoS Pathogens (2008) 4(12): e1000251 doi:10.1371/journal.ppat.1000251

(h) Research grants held by supervisor
Research grants held in the last five years and any key prior grants (list the most recent first.
Please state the name of the awarding body, title of project, amounts awarded and start and end dates of support. For all current grants, indicate the number of hours per week that are spent on each project.

Grants from other funding agencies
MRC: Structural Studies on Flu polymerase £1,500,000 2007-2009. 5hrs
EU: VIZIER Structural Genomics of Viral Enzymes Involved in Replication Oxford component: 1,215404 euros. 2005-2009. 5hrs

Grants from the Trust

Q11 DETAILS OF SPONSOR

(a) Surname: harvey Forenames: paul

(b) Title of current post: Head of Department, Professor of Zoology

Date of appointment: 01/10/1985

Expected date of termination: 30/09/2012

(c) With whom do you have your contract of employment?

University of Oxford

(d) Source of personal salary support (If ‘Other’, please specify):

Please also be specific is salary is funded from more than one source.

HEFC

Q11 DETAILS OF SPONSOR

(a) Surname: Stuart Forenames: David Ian

(b) Title of current post: MRC Research Professor, Oxford.

Date of appointment: 01/10/1995

Expected date of termination: 01/12/2018

(c) With whom do you have your contract of employment?
University of Oxford

(d) Source of personal salary support (If ‘Other’, please specify):

Please also be specific is salary is funded from more than one source.

Medical Research Council

I am employed 50% as Life Science Director by Diamond (since April 2008) and 50% MRC Porfessor in Oxford.

Q12 RESEARCH QUESTION

(a) What is your research question? (no more than 100 words)

What are the characteristic structural and functional properties of proteins created de novo (i.e. not by duplication of existing genes), in viruses and in bacteria?

Answering this question will allow me to address two broader evolutionary issues:

- Why is there a relatively small number of protein structural folds in nature? Are some folds favoured for functional or physical reasons, or are most proteins descended by duplication from a limited number of ancestral proteins?
- What is the role of de novo protein creation on the cross-species transmission (emergence) and pathogenicity of viruses and of bacteria?

(b) Why is it important?
(no more than 250 words)

Recent studies have shown that many more proteins are created de novo than thought previously. Such proteins tend to be young and allow us to study the evolution of protein structure and function. This is important for the following reasons:

(i) Most viral proteins created de novo that I identified in an earlier study play a role in pathogenicity. Understanding their properties can thus help us understand how they contribute to viral emergence and virulence. Likewise, de novo proteins of commensal and pathogenic bacteria are suspected to play a role in adaptation to their host.

(ii) Do proteins created de novo have novel structures, or do they adopt typical folds that are seen in old proteins? Does this change through evolutionary time?
Finding that most de novo proteins have known folds would strongly suggest that the number of folds is physically limited. On the other hand, finding that de novo proteins adopt previously unknown folds would require us to develop testable hypotheses for why the main protein families adopt so few folds. It might even suggest that we are underestimating the number of protein folds because of our limited knowledge of the structure of de novo proteins.

(iii) One of the key problems of current bioinformatics research is ab initio protein structure prediction. For this we need to know the widest possible library of folds.

Knowing the properties of these proteins will help the emerging field of de novo protein creation to coalesce around a set of fundamental observations.

Q13 SUMMARY OF PROPOSED RESEARCH INCLUDING KEY GOALS

(a) For scientifically qualified assessors: (no more than 200 words)

Novel proteins are thought to be created mostly through gene duplication. However, recent studies showed that de novo protein creation occurs at an unexpectedly high rate. I have shown that a particular subset of de novo proteins, those encoded by viral overlapping genes, are abundant and relatively easy to identify. These proteins have unusual sequence and structure properties (being unstructured or having previously unobserved structural folds), and specific functions (usually associated with viral pathogenicity).

In collaboration with two experienced investigators with complementary skills, I will collect and curate a much larger dataset of several hundreds de novo proteins from viruses (mainly) and from bacteria. I will study their sequence, evolution, function, and structure, through bioinformatics and experimental approaches. In particular I aim to solve about ten of their 3D structures. This will provide a new, experimental approach to understand the evolution of protein structure and possibly challenge the belief that nature creates proteins only according to a limited number of folds.

This research will also improve our scant knowledge of viral accessory proteins, which are often created de novo, and may uncover features associated with virus emergence.

Each step of the project will generate findings that can be published independently.

(b) For lay readers: (no more than 200 words)

Evolution is mainly thought to proceed by "tinkering", with new proteins being created from "re-used" pieces of existing proteins. However, recent studies, including mine, have shown that a significant number of proteins are in fact created "from scratch".

Very little is known about these novel proteins. I will collect many new ones in viruses and in bacteria, using a "trick" by focusing on proteins encoded by particular segments of DNA called "overlapping genes". I will study their properties, including their three-dimensional structure. At present, it is thought that nature creates proteins only according to a limited number of shapes. However, my previous work suggests that nature might well create novel proteins in all shapes. This work is thus a new approach to an old, unsolved problem. Besides, these novel proteins might allow viruses to infect new hosts or to become more pathogenic. Thus, studying them might hep us to understand how viruses move from one host to another and become more virulent.

Understanding these novel proteins and their evolution will lead onto new questions. For instance, do complex organisms "re-use" proteins more than simple ones such as viruses? If so, why?

Q14 DETAILS OF RESEARCH PROJECT

Detail (a) Aims of the project, (b) Work which has led up to the project, (c) Experimental design and methods to be used in investigating this problem. Full details of the study design for experiments (humans & animals) must be provided. This should include power calculations, sample size justification and, where appropriate, case definition & inclusion/exclusion criteria. Please refer to guidelines. (d) Timetable and milestones, if appropriate. **For clinical trials, refer to guidelines**.

No more than 3,500 words should be used to describe the research project.

Graphs, figures and supporting unpublished data may be embedded in the text or included as an appendix or appendices. These additional data must not exceed the equivalent of 5 A4 pages in length.

(a) AIMS OF PROJECT

I aim to determine the sequence, structural, and functional properties of proteins having been created *de novo* (i.e. not by duplication). I will use a combination of bioinformatics and biochemical approaches. Identifying proteins created *de novo* is very difficult in general (Long et al. 2003) because it requires a number of well-annotated closely related genomes (Toll-Riera et al. 2008, Zhou et al. 2008). To circumvent this problem, I will use a very specific category of *de novo* proteins: those encoded by overlapping genes. Indeed, overlapping genes are created *de novo* by mutations within an ancestral coding sequence that leads to the expression of a novel protein in another reading frame, a process called "overprinting" (Keese and Gibbs 1992) (Fig.1). The identification of the novel protein and of the ancestral one in each overlap can be done by relatively simple evolutionary analyses (see "experimental design"). This approach will thus allow the identification of an unprecedented number of proteins created *de novo*.

The specific aims of the project are:

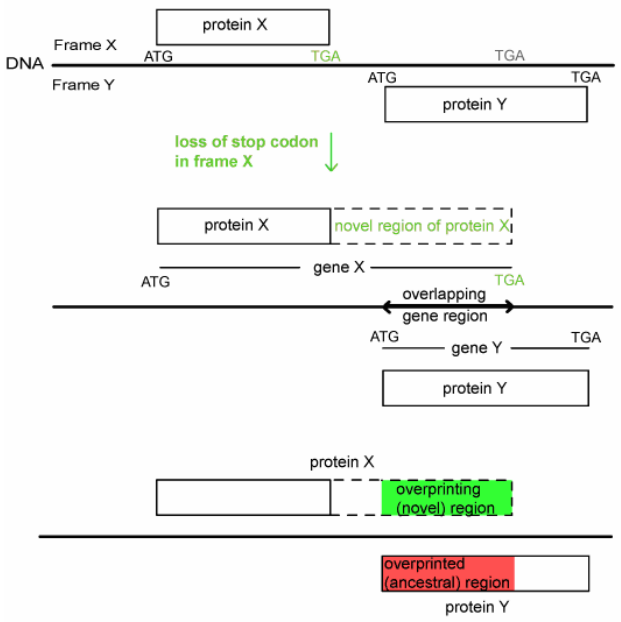
- build a curated bioinformatics dataset of about 200 viral and 50 bacterial proteins or protein regions created *de novo* by overprinting. The dataset will include their sequence, their annotated functions and relevant features and be freely available to the research community.
- study their sequence properties and identify their most common known functions
- express and purify those predicted to be ordered, and elucidate the function of relevant ones in collaboration.
- solve the 3D structure of at least 5 viral and 5 bacterial *de novo* proteins.

Potential follow-up studies

As with any emerging field, opportunities will abound at the end of the project. Here are possible collaborations that could stem from it:

- carry out the same type of analysis on eukaryotic *de novo* proteins, on which almost nothing is known (Toll-Riera et al. 2008).
- solve the structure of *de novo* membrane proteins, which would be of considerable interest given our scant knowledge of membrane proteins in general.
- study organisms that do not use gene duplication to create novelty. For instance, in the fungus *Neurospora Crassa*, duplication is prevented by a mechanism called "repeat-induced point mutation" (Brookfield 2003), and thus *de novo* protein creation might play a major role (Braun et al 2000).

Figure 1 - Creation of a novel protein region by overprinting



Top: a DNA sequence encodes 2 proteins in different reading frames. Notice the potential, unused stop codon downstream of protein X. Middle: a mutation abolishes the stop codon of protein X, causing its elongation ("overprinting") until the pre-existing stop codon, resulting in a gene overlap. Below: the overlap encodes an overprinted (ancestral) protein region, in red, and an overprinting (novel) one, in green.

(b) BACKGROUND: WORK WHICH HAS LED TO THE PROJECT

In 2002, at the end of my PhD, I made an intriguing observation: most protein domains encoded by overlapping genes of measles virus were disordered (unstructured) (Karlin et al. 2003). Since then I have worked in science communication but in my spare time I carried out and coordinated a work, described below, which confirmed the initial intuition and evolved into the more general question of the *de novo* creation of proteins by viruses (Rancurel et al 2009; the article is undergoing minor revisions for J Virol. A figure summing up our main findings is enclosed as additional file 1).

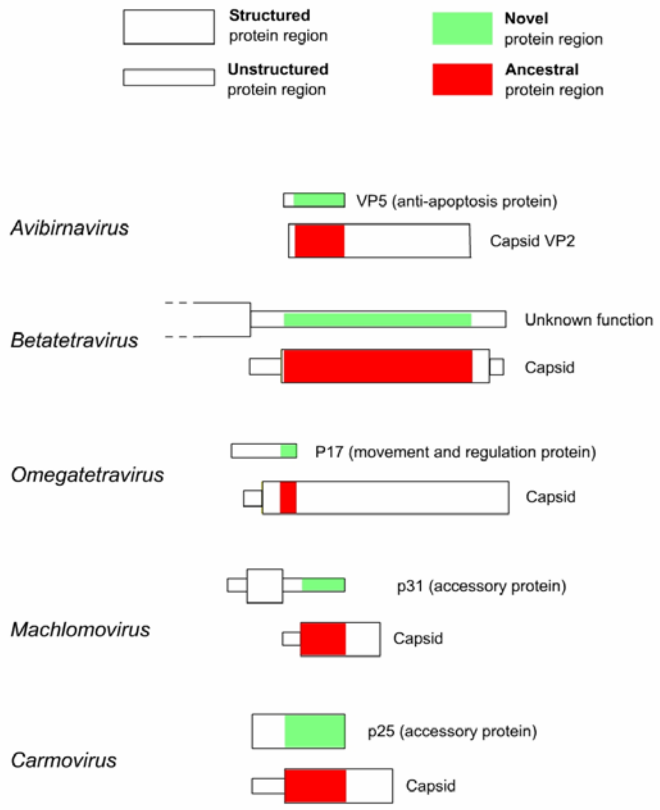
In parallel, Robert Belsaw and Jonathan Grimes made complementary observations on the evolution of overlapping genes, also described below (Meier et al. 2006, Belshaw et al. 2007). The present project is a natural extension of my previous work and of our ideas, but focuses on *de novo* proteins that are ordered rather than disordered.

1.Viral proteins created *de novo* by overprinting have unusual sequence properties and are often involved in viral pathogenicity

In my previous study (Rancurel et al. 2009), I analyzed the sequences of the protein products of manually curated overlapping genes from 43 genera of unspliced RNA viruses infecting eukaryotes. I found that overlapping proteins have a sequence composition biased towards disorder-promoting amino acids and are predicted to contain significantly more structural disorder than non-overlapping proteins, thus confirming my initial observation.

By analysing the phylogenetic distribution of overlapping proteins, we could confirm that 17 of these had been created *de novo* and studied them individually (Additional File 1). Almost all were accessory proteins that play a crucial role in viral pathogenicity or spread, rather than proteins central to viral replication or structure (Fig.2). 60% were predicted to be fully disordered and had a highly unusual sequence composition. 40% were predicted to be ordered, and whenever their 3D structure had been solved, it corresponded to a fold previously unobserved.

Figure 2 - Overprinting of viral capsid protein sequences by proteins created *de novo*



A particular set of *de novo* proteins (in green) from my previous study: those created by overprinting the sequences of ancestral capsid genes (in red). Conventions are the same as in Fig. 1.

The capsid genes are homologous. Note that:
- most novel protein regions have very different functional or structural properties, despite being created from homologous DNA sequences.
- their functions are mainly associated with **viral pathogenesis or spread**.

The protein p25 (bottom) is part of the **preliminary dataset** of proteins suitable for X-ray crystallography (additional File 3; see text)

2.Parallel observations by Jonathan Grimes on the structure of a *de novo* protein

The structure of one *de novo* protein, NSP9 from the SARS *coronavirus*, created by overprinting the viral nucleoprotein, has been solved by Jonathan Grimes' group (Meier et al. 2006). It contained an unusual amount of structural disorder, while ordered parts adopted a previously unknown structural fold. J. Grimes made the hypothesis that both features were consequences of the *de novo* creation of NSP9, and independently suggested that solving the 3D structures of *de novo* proteins might enhance our understanding of the evolution of protein structure.

3.Large-scale evolutionary analysis of viral overlapping genes by Robert Belshaw

In parallel, Robert Belshaw carried out an automated analysis of overlapping genes from RNA viruses (Belshaw et al. 2007). Its dataset was larger than mine, including spliced viruses and bacteriophages, but was not curated. He found that larger viral genomes tended to have less overlaps and proposed a testable model for the evolution of gene overlaps.

4.*De novo* protein creation is not limited to viruses and accounts for a portion of "orphan" proteins in other organisms

"Orphan", or "taxonomically restricted" proteins are proteins having no detectable homologue in other organisms (Wilson et al. 2005). Their widespread occurrence in every sequenced genome remains unexplained. Two systematic studies on *de novo* protein creation in eukaryotes were published recently. They focused on creation from non-coding sequences, as opposed to overprinting. The studies indicate that *de novo* protein creation is not restricted to overlapping genes and occurs at an unexpected rate, having generated between 5% and 20% of orphan proteins of primates (Toll-Riera et al. 2008), and about 12% in the genus *Drosophila* (Zhou et al. 2008). Thus this research topic is also highly relevant to eukaryotes.

No such systematic study of *de novo* bacterial proteins exist but studies of orphan proteins from pathogenic and commensal bacteria suggested that many had been created *de novo* and played a role in adaptation to their hosts (Deckers et al. 2004, van Passel et al. 2008).

Rationale for the present project: combining our strengths

Robert Belshaw, Jonathan Grimes and I bring complementary approaches to the project. R. Belshaw proposed a model of evolution of overlapping genes, making interesting predictions, but could not test them because, not being an expert in protein bioinformatics he could not determine which proteins were novel or ancestral, or elucidate their domain organisation. J. Grimes was willing to solve the structure of *de novo* proteins and had the necessary expertise, but could invest the time required to curate overlapping genes. I could not draw reliable conclusions concerning R. Belshaw's model of evolution owing to the relatively small size of my dataset of *de novo* proteins, and could not study their functions because I had no access to virology labs. Therefore we have decided to combine our strengths and collect a large number of *de novo* proteins encoded by overlapping genes to study their evolution, structure and function.

(c) EXPERIMENTAL DESIGN AND METHODS.

This project is composed of two parts: an evolutionary, bioinformatics part in Robert Belshaw's lab to build the dataset of *de novo* proteins, and an experimental part (structural and functional) in Jonathan Grimes' lab.

I will be based, and have office and computer facilities, at R. Belshaw's lab, but in years 2 and 3 will spend 4 days a week at J. Grimes' lab, which will provide all the necessary laboratory facilities and is only a short bicycle ride away.

In year 1, in parallel to the first part, I will carry out a pilot of the second part in Jonathan Grimes's lab using a small, unpublished dataset of *de novo* proteins suitable for structure resolution (additional file 3). This will allow ironing out any teething problems before the full scale study.

We have taken every precaution to ensure that each step will bring results exploitable by the community and publishable on their own. These steps and the choice of our model systems are described below.

0 - CHOICE OF MODEL SYSTEMS

1.Viral overlapping genes

I chose to work on proteins created *de novo* in viral overlapping genes for several reasons:

- each overlap encodes one protein created *de novo* by overprinting, relatively easy to identify compared to proteins having arisen *de novo* from non-coding sequences
- viruses contain numerous long overlapping genes and so I will be able to gather numerous *de novo* proteins or protein domains.
- in particular, Andrew Firth's team, using gene prediction software he developed, recently discovered 5 overlooked overlapping genes, including several ones that are essential for the virus (Chung et al. 2008). Thus overlapping genes are probably even more important than previously thought.
- because of their small size, viral genomes are typically much better annotated than eukaryotic genomes, which facilitates curation. This also means that a function will have been assigned experimentally to the majority of *de novo* viral proteins, contrary to eukaryotic ones. Thus in viruses we can readily have information about the functional impact of *de novo* protein creation.
- in particular, my preliminary analysis (Rancurel et al. 2009) and others (Young et al. 2000, Hout et al. 2004, Li and Ding 2006) suggest that *de novo* proteins play a crucial role in viral pathogenicity and spread, at least in mammalian and plant RNA viruses. Yet these crucial accessory proteins have been the focus of very little attention compared to enzymes or structural proteins (Fogg et al 2006). Thus this model system is of particular relevance to human, animal and plant health.

We will study only viruses infecting eukaryotes since for others (bacteriophages), curation would be too difficult, from my experience.

2.Bacterial overlapping genes

Almost nothing is known about *de novo* bacterial proteins, despite their suspected role in host or environmental adaptation (see above). I will carry out the same study as on viral proteins on about 50 *de novo* bacterial proteins encoded by overlapping genes. Since they are much easier to crystallise than viral proteins (Structural Genomics Consortium et al. 2008), I can reasonably expect to solve the structure of at least 5 *de novo* bacterial proteins.

A potential drawback of working with *de novo* proteins encoded by overlapping genes is that their sequence composition can be influenced by the fact that they overlap another reading frame, *i.e.* we are not certain that they are representative of typical proteins created *de novo* from non-coding sequences. I am aware of this drawback and will be careful not to extrapolate my conclusions unreasonably.

1 – BIOINFORMATICS PART CARRIED OUT IN ROBERT BELSHAW'S TEAM

1.Collection and curation of a dataset of viral and bacterial overlapping genes– year 1

Because of essential differences in viral and bacterial genomes, the collection process of both datasets will be different:

a. Collection of long (>90nt) viral overlapping genes

I will collect gene overlaps in all completely sequenced eukaryotic viruses, keeping only those whose existence has been proven experimentally. I will study only overlaps >90 nucleotides, corresponding to 30aa, for two reasons:
1) shorter regions are unlikely to fold by themselves (Stricher et al. 2006) and thus expected to have a lesser structural impact.
2) the reliability of disorder prediction increases with length (Obradovic et al. 2003).

I will proceed essentially as in my previous work (Rancurel et al. 2009) but collect a much larger dataset, including spliced RNA viruses and all DNA viruses. I will use perl scripts (also written by Robert Belshaw's team for his previous study) to collect all (>1500) genome sequence files from the NCBI viral database (Bao et al. 2004), and parse them to detect overlaps >90nt.

I will then curate overlaps by looking for experimental evidence that both encoded proteins are expressed, using two complementary methods:
1) if the proteins are referenced in the high quality, curated protein database SwissProt, look for annotation of experimental evidence that they exist (Bairoch et al. 2005).
2) extensive bibliographical analyses for proteins not present in SwissProt.

b .Collection of bacterial overlapping genes
Overlapping genes are frequent in bacteria but typically very short (a few nucleotides) (Fukuda et al. 2003). Because of their larger size, bacterial genomes contain numerous long (>90nt), spurious overlapping reading frames (Pallejà et al 2008), making manual curation impossible. However, I have assembled, by bibliographical searches, an unpublished collection of about 20 *bona fide* bacterial gene overlaps longer than 90nt (additional file 2). A good part are not referenced in databases, which suggests that they are more common that previously thought, but poorly referenced. In all, I expect to find at least 50 *bona fide* bacterial overlaps >90nt through extensive bibliographical searches.

My previous experience in curating overlapping genes is a strong asset, since few people are prepared to invest the necessary time, resulting in lower quality datasets.

2.Identification of the proteins created *de novo* by evolutionary analyses – year 1
As a reminder, each overlap encodes one ancestral protein and one protein created *de novo* by overprinting it (see Aims and Fig.1). I will identify both types using a combination of two approaches (see Fig.3):

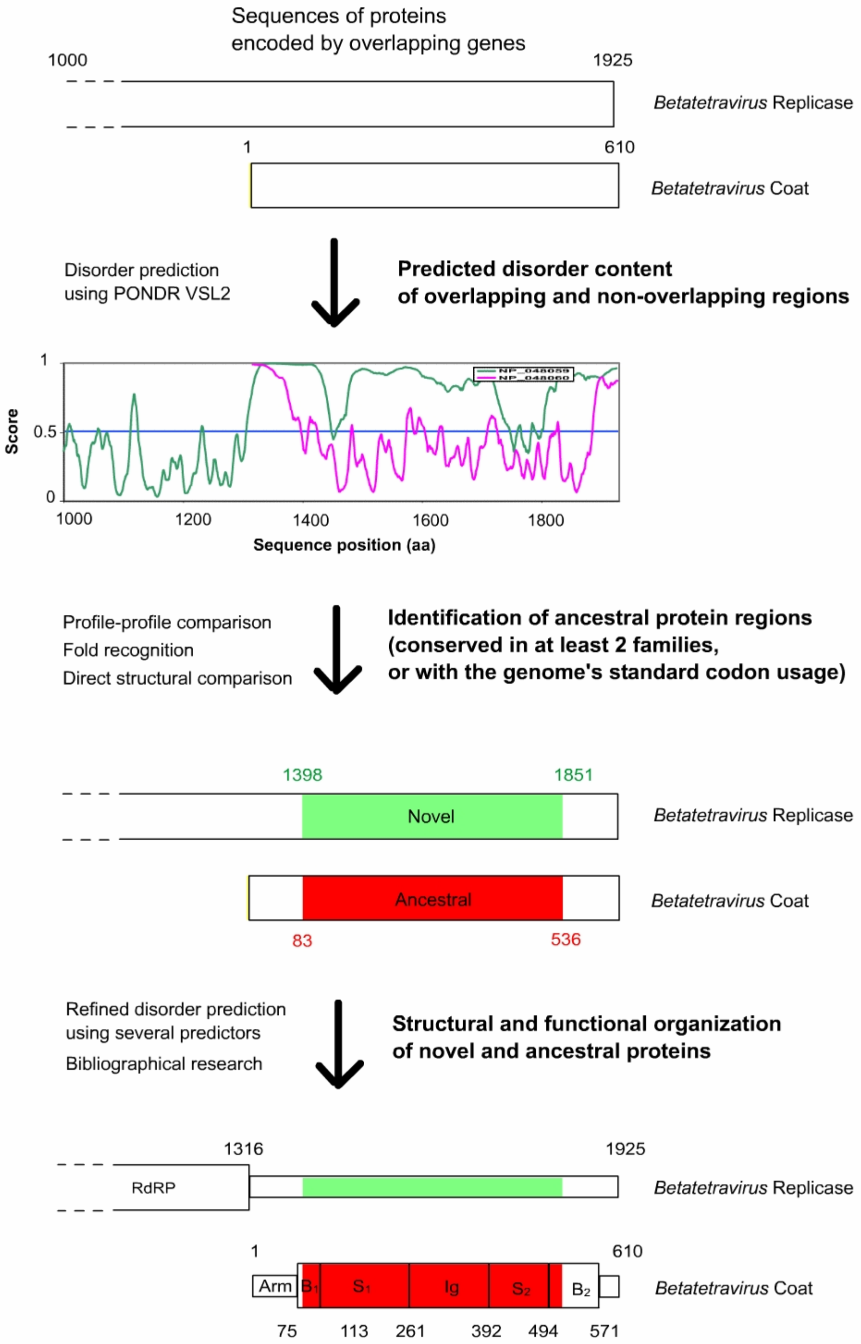
- I will analyse **phylogenetic conservation** of each protein, as in my previous study. Proteins conserved in at least 2 viral or bacterial taxonomical families are deemed ancestral. Given the fast rate of evolution of viruses and bacteria this is a very conservative criterion. I will assess sequence and structure similarity using respectively sequence profile-profile comparison and fold recognition or direct structural comparison. In my previous study I could identify with confidence the ancestral protein (and thus the *de novo* protein overlapping it) in 40% of overlaps.

- Robert Belshaw will analyse the **codon usage** of overlapping reading frames. This method has the advantage of not requiring the existence of any homolog and is thus complementary to the previous one. In an earlier study (Pavesi 1997), most (86%) overlaps encoded one protein having the "standard" codon usage of the genome (and thus deemed ancestral) and one protein with a very different codon usage (deemed novel).

Our previous experience (Rancurel et al. 2009) suggests good agreement between both methods (we compared the results of our phylogenetic analysis to those of codon usage analysis reported in the literature). Of course we will discard proteins for which there is no agreement.

From the above figures, we thus expect to determine the ancestral and novel protein for **almost all overlaps of the dataset**.

Figure 3 - Workflow for structural and functional bioinformatics analysis: ex. of the betatetravirus replicase/capsid overlap


Sequences of proteins
encoded by overlapping genes
1000
1925
Betatetravirus Replicase
1
610
Betatetravirus Coat
Disorder prediction
using PONDR VSL2
Predicted disorder content
of overlapping and non-overlapping regions
Score
1
0.5
0
NP_048059
NP_048060
1000
1200
1400
1600
1800
Sequence position (aa)
Profile-profile comparison
Fold recognition
Direct structural comparison
Identification of ancestral protein regions
(conserved in at least 2 families,
or with the genome's standard codon usage)
1398
1851
Novel
Betatetravirus Replicase
Ancestral
Betatetravirus Coat
83
536
Refined disorder prediction
using several predictors
Bibliographical research
Structural and functional organization
of novel and ancestral proteins
1316
1925
RdRP
Betatetravirus Replicase
1
610
Arm
B1
S1
Ig
S2
B2
Betatetravirus Coat
75
113
261
392
494
571

Conventions are the same as in Fig. 1. Second panel, superimposed PONDR disorder prediction for the betatetravirus capsid (magenta) and replicase (green). Regions with a score above 0.5 are predicted disordered.
Third panel, predictions of the boundaries of ancestral and novel regions of the overlapping proteins. Below: refined structural and functional analysis. Domain names were obtained from the literature.

3.Bioinformatics analysis of novel and ancestral proteins: sequence features, functions, evolution – year 2 and 3

- Using bioinformatics software, I will analyse the sequence composition of each protein, its predicted structural and functional features (ordered or disordered regions, transmembrane segments, metal-binding motifs, etc.) and domain organisation, as in my previous study (Fig.3).
- I will test Robert Belshaw's model of creation and evolution of overlapping genes (Belshaw et al. 2007) by studying their mechanism of creation (whether by a mutation abolishing a stop codon, or giving rise to a new start codon, or to a splicing event...) and by estimating the relative age ranges of *de novo* proteins using their phylogenetic distribution.
- The large size of the dataset will allow quantitative identification of features specific to *de novo* proteins (for instance enrichment in a particular function or in a metal-binding motifs compared to ancestral proteins), using the software PROMPT (http://webclu.bio.wzw.tum.de/prompt) (Schmidt and Frishman 2006).
- In particular, I will try to identify patterns of *de novo* proteins associated with the pathogenicity and choice of host of viruses.

4.Selection of suitable targets for structure determination

I will select from the previous step the *de novo* proteins or domains predicted to be ordered and non-membranar. I will include homologous proteins from related viruses or bacteria as targets, since it enhances considerably the success of structural genomics projects (Jaroszewski et al 2008). In my previous work most *de novo* proteins had at least one homologue and thus inclusion of homologues should be easy.

In cases where a domain and not a full-length protein is selected for expression, we will construct 3 different constructs per domain, to maximise chances of soluble expression (Peti and Page 2007).

From my previous work (Rancurel et al 2009) I expect to have about 50 viral and 15 bacterial protein targets suitable for X-ray crystallography. Including the homologues and the different constructs, this amounts to about 200-250 targets. With current high-throughput structural genomics techniques, this is achievable in 2 years (see timetable) by a single person (with the help of Jonathan Grime's team of course and given due training – see Training section).

I will handle the process as described below.

2 – EXPERIMENTAL PART CARRIED OUT IN JONATHAN GRIMES' TEAM

1.Small scale pilot – year 1

In my earlier work I identified about 10 viral and bacterial proteins created *de novo* suitable for crystallization. I will immediately start cloning, expressing and purifying them and their homologues, in parallel with the dataset-building stage of the large scale study above. This phase is very important since it will allow me to iron out any teething problems in the biochemical part of the project.

2.High-throughput cloning, expression and purification in *E coli* of the *de novo* target proteins – year 2

I will use the world-class Oxford Protein Production Facility (OPPF) high-throughput pipeline (Berrow et al 2007), using the general strategy recently outlined by a wide variety of structural genomics consortia (Structural Genomics Consortium et al. 2008), adapted for viral proteins from the OPP experience. *E coli* will be my first attempt for expression because the experience of OPPF consistently showed it is a far cheaper and faster alternative to other organisms. However, if *E coli* expression fails to bring enough proteins to purification, I will use baculovirus expression, for which Jonathan Grimes' lab, the Division of Structural Biology (STRUBI) has all the required expertise.

Cloning, Expression and Purification

Using a ligation-independent method, I will clone the synthetic DNA templates of target proteins, optimised for expression in *E coli,* in OPPF expression vectors (Berrow et al. 2007), comprising cleavable N-term and C-term His_6 fusion tags.

I will express the proteins in small-scale 96 well-plates and then optimise expression conditions. I will purify soluble proteins using parallelized AKTA Xpress immobilized metal affinity chromatography (IMAC) followed by gel-filtration.

3.Functional analysis of purified *de novo* proteins – year 2 and 3

We will establish collaborations with virology and microbiology labs for functional analysis of relevant purified proteins (J. Grimes was part, among others, of the EC funded VIZIER project (www.vizier-europe.org) and thus has a dedicated network of collaborations from virologists). We will select purified proteins that either come from viruses (or bacteria) with important pathogenic roles in humans, animals or plants, or that have a previously unknown activity.

4.Crystallization and resolution of the 3D structure of *de novo* proteins by X-ray crystallography – year 2 and 3

Crystallisation

I will benefit from state-of-the-art techniques with proven efficiency to crystallise viral proteins, for instance nanolitre-scale crystallisation (Walter TS et al. 2005) and automated crystallisation plate imaging systems at 4° and 21°C (Mayo CJ et al. 2005). I will also use techniques to optimise the protein sample for crystallisation, *eg* thermofluor (which directly measures the melting temperature of the protein) and surface entropy reduction by chemical methylation (Walter et al. 2006) (Geerlof et al. 2006).

In-house crystallographic analysis

STRUBI X-ray diffraction facilities enable crystal screening. Because this final step of structure resolution, is very time consuming, I will benefit from the help of other crystallographers from J. Grime's lab.

Synchrotron access and phasing

Virtually all the final diffraction datasets will be collected using synchrotron radiation. We will use SeMet labelling followed by MAD or SAD analysis at the ESRF BM14 beamline and at Diamond, once available. STRUBI has an established track record in high-level incorporation and successful structure determination for SeMet proteins (Aricescu et al. 2006) and I will benefit from their help.

(D) TIMETABLE AND MILESTONES

YEAR 1

Small scale phase

Expression, purification and crystallization of about 10 suitable *de novo* proteins identified from my previous study. This phase, done in parallel with the bioinformatics phase of the large scale study, is very important since it will allow me to iron out any teething problems before embarking in the biochemical part.

Large scale study

Milestone1. Collection and curation of a dataset of 150 viral and 50 bacterial overlapping genes.

Milestone2. Identification of the protein created *de novo* in each pair of gene overlaps by evolutionary analyses
Milestone3 Selection of about 60 suitable targets for structure resolution

YEAR 2

Cloning and expression of the target proteins.
Milestone4. Purification of the target proteins.
Initial crystallization trials.
Milestone5. Bioinformatics analysis of overlapping proteins: sequence composition, annotated functions.

YEAR 3

Refinement of crystallization.
Milestone6. Functional analysis of relevant purified proteins in collaboration with external labs.
Milestone7. Resolution of the 3D structure of 10 *de novo* proteins or protein domains.

Q15 REFERENCES (Research project)
Please give citation in full, including title of paper and all authors.

Altman RB (2004) Editorial: Building successful biological databases. Brief Bioinform 5: 4-5.

Aricescu AR, Assenberg R, Bill RM, Busso D, Chang VT, Davis SJ, Dubrovsky A, Gustaffson L, Hedfalk K, Heinemann U, Jones IM, Ksiazek D, Lang C, Maskos K, Messerschmidt A, Maciera S, Peleg Y, Perrakis A, Poterszman A, Schneider G, Sixma TK, Sussman JL, Sutton G, Tarbouriech N, Zeev-Ben-Mordehai T, Jones EY (2006). Eukaryotic expression: developments for structural proteomics. Acta Crystallogr D Biol Crystallogr 62 (10) 1114-24

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33(Database issue): D154-159.

Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S et al. (2004) National center for biotechnology information viral genomes project. J Virol 78(14): 7291-7298.

Belshaw R, Pybus OG, Rambaut A (2007) The evolution of genome compression and genomic novelty in RNA viruses. Genome Res 17(10): 1496-1504.

Belshaw, R., de Oliviera, T., Markowitz, S. & Rambaut, A. (2009). The RNA Virus Database. Nucleic Acids Research. 37: D431-D435.

Berrow NS, Alderton D, Sainsbury S, Nettleship J, Assenberg R, Rahman N, Stuart DI, Owens RJ (2007) A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucleic Acids Res 35(6):e45.

Birney E, Clamp M (2004) Biological database design and implementation. Brief Bioinform Mar;5(1):31-8.

Braun EL, Halpern AL, Nelson MA, Natvig DO (2000) Large-Scale Comparison of Fungal Sequence Information: Mechanisms of Innovation in Neurospora crassa and Gene Loss in Saccharomyces cerevisiae. Genome Res 2000 10: 416-430.

Brookfield JF (2003)Genome Sequencing: The Ripping Yarn of The Frozen Genome. Current Biology, Vol. 13, R552â€“R553,

Chung BY, Miller WA, Atkins JF, Firth AE (2008) An overlapping essential gene in the Potyviridae. Proc Natl Acad Sci U S A. 105(15):5897-902

Deckers D, Masschalck B, Aertsen A, Callewaert L, Van Tiggelen CG, Atanassova M, Michiels CW (2004) Periplasmic lysozyme inhibitor contributes to lysozyme resistance in Escherichia coli. Cell Mol Life Sci 61(10):1229-37.

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V et al. (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34(Database issue): D247-251.

Fogg MJ, Alzari P, Bahar M, Bertini I, Betton JM, Burmeister WP, Cambillau C, Canard B, Corrondo MA, Coll M, Daenke S, Dym O, Egloff MP, Enguita FJ, Geerlof A, Haouz A, Jones TA, Ma Q, Manicka SN, Migliardi M, Nordlund P, Owens RJ, Peleg Y, Schneider G, Schnell R, Stuart DI, Tarbouriech N, Unge T, Wilkinson AJ, Wilmanns M, Wilson KS, Zimhony O, Grimes JM (2006) Application of the use of high-throughput technologies to the determination of protein structures of bacterial and viral pathogens. Acta Crystallogr D Biol Crystallogr. Oct;62(Pt 10):1196-207

Fukuda Y, Nakayama Y, Tomita M (2003) On dynamics of overlapping genes in bacterial genomes. Gene 323: 181-187.

Galperin MY (2008) The Molecular Biology Database Collection: 2008 update. Nucleic Acids Res. Jan;36(Database issue):D2-4

Geerlof A, Brown J, Coutard B, Egloff MP, Enguita FJ, Fogg MJ, Gilbert RJ, Groves MR, Haouz A, Nettleship JE, Nordlund P, Owens RJ, Ruff M, Sainsbury S, Svergun DI, Wilmanns M (2006) The impact of protein characterization in structural proteomics. Acta crystallogr D Biol Crystallogr 62: 1125-36.

Hardin C, Pogorelov TV, Luthey-Schulten Z Ab initio protein structure prediction (2002). Curr Opin Struct Biol. Apr;12(2):176-81.

Hout DR, Mulcahy ER, Pacyniak E, Gomez LM, Gomez ML, Stephens EB (2004). Vpu: A multifunctional protein that enhances the pathogenesis of human immunodeficiency virus type 1. Curr. HIV Res. 2: 255-270.

Jaroszewski L, Slabinski L, Wooley J, Deacon AM, Scott AL, Wilson IA, Godzik A (2008). Genome Pool Strategy for Structural Coverage of Protein Families. Structure 16, 1659â€“1667.

Karlin D, Ferron F, Canard B, Longhi S (2003) Structural disorder and modular organization in Paramyxovirinae N and P. J Gen Virol 84(Pt 12): 3239-3252.

Keese PK, Gibbs A (1992) Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci U S A 89(20): 9489-9493.

Li F, Ding SW (2006) Virus counterdefense: diverse strategies for evading the RNA-silencing immunity. Annu Rev Microbiol 60: 503-531.

Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4(11): 865-875.

Mayo CJ, Diprose JM, Walter TS, Berry IM, Wilson RJ, Owens RJ, Jones EY, Harlos K, Stuart DI, Esnouf RM (2005) Benefits of automated crystallization plate tracking, imaging and analysis. Structure 13(2):175-82.

Meier C, Aricescu AR, Assenberg R, Aplin RT, Gilbert RJ et al. (2006) The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus. Structure 14(7): 1157-1165.

Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ et al. (2003) Predicting intrinsic disorder from amino acid sequence. Proteins 53 Suppl 6: 566-572.

PallejÃ A, Harrington ED, Bork P (2008) Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics Jul 15;9:335.

Pavesi A, De Iaco B, Granero MI, Porati A (1997) On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. J Mol Evol 44(6): 625-631.

Peti W, Page R (2007) Strategies to maximise heterologous protein epression in Escherichia coli with minimal cost. Protein Expr Purif 51:1-10

Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. In revision for J Virol (May 2009).

Schmidt T, Frishman D PROMPT: a protein mapping and comparison tool (2006). BMC Bioinformatics 7:331-345

Stricher F, Martin L, Vita C (2006) Design of miniproteins by the transfer of active sites onto small-size scaffolds. Methods Mol Biol 340: 113-149.

Structural Genomics Consortium, Architecture et Fonction des MacromolÃ©cules Biologiques, Berkeley Structural Genomics Center, China Structural Genomics Consortium, Integrated Center for Structure and Function Innovation, Israel Structural Proteomics Center, Joint Center for Structural Genomics, Midwest Center for Structural Genomics, New York Structural GenomiX Research Center for Structural Genomics, Northeast Structural Genomics Consortium, Oxford Protein Production Facility, Protein Sample Production Facility, Max DelbrÃ¼ck Center for Molecular Medicine, RIKEN Structural Genomics Proteomics Initiative, SPINE2-Complexes (2008) Protein production and purification. Nature Methods 5(2):135-46

Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet.;38:615-43

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, AlbÃ MM (2008) Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol. sDec 8.

van Passel MW, Marri PR, Ochman H (2008) The emergence and fate of horizontally acquired genes in Escherichia coli. PLoS Comput Biol 4(4)

Walter TS, Diprose JM, Mayo CJ, Siebold C, Pickford MG, Carter L, Sutton GC, Berrow NS, Brown J, Berry IM, Stewart-Jones GB, Grimes JM, Stammers DK, Esnouf RM, Jones EY, Owens RJ, Stuart DI, Harlos K (2005) A procedure for setting up high-throughput nanolitre crystallization experiments. Crystallization workflow for initial screening, automated storage, imaging and optimization. Acta Crystallogr D Biol Crystallogr 61: 651-7.

Walter TS, Meier C, Assenberg R, Au KF, Ren J, Verma A, Nettleship JE, Owens RJ, Stuart DI, Grimes JI. Lysine methylation as a routine rescue strategy for protein crystallization. Structure 14(11):1617-22

Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D (2005) Orphans as taxonomically restricted and ecologically important genes. Microbiology 151(Pt 8):2499-501

Young I, Wang I, and Roof WD (2000). Phages will out: strategies of host cell lysis. Trends. Microbiol. 8: 120-128.

Zeldovich KB, Chen P, Shakhnovich BE, Shakhnovich EI (2007) A first-principles model of early evolution: emergence of gene families, species, and preferred protein folds. PLoS Comput Biol. Jul;3(7):e139.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W (2008). On the origin of new genes in Drosophila. Genome Res Sep;18(9):1446-55.

Q16 RESEARCH TRAINING

Please outline the research training proposed including dates and locations of training. (No more than 700 words).

I am very eager to take the opportunity of a Wellcome Trust career re-entry grant to undergo training, especially since I was myself a vocational trainer during my career break from research and thus can appreciate the value of training.
Below is the training I will gain from the Division of Mathematical, Physics and Life Sciences of the University of Oxford (www.ox.ac.uk/divisions/mpls.html). I will of course look for other opportunities to strengthen the areas where I lack (in particular training offered by the Wellcome Trust), according to the development of the project.

- Bioinformatics and evolutionary biology

I am experienced with this part of the project, ie the gathering and curating of the viral dataset and the bioinformatics analysis. However, the larger size of the dataset will allow new quantitative analyses that require a better mastery of statistics, and I will thus undergo training in STATISTICAL ANALYSIS OF BIOLOGICAL DATASETS. The field and the tools of protein bioinformatics and of its evolutionary aspects are fast moving. Therefore, I will also undergo a more general training in PROTEIN EVOLUTION, SEQUENCE, STRUCTURE AND SYSTEMS.

- Structural genomics

I am very experienced in bringing proteins to expression and purification on a small-scale, and have kept my bench skills up to date during my break from research by teaching vocational training in molecular biology and biochemistry. However, I need to learn large-scale protein expression, purification, and crystallization.
I will receive a thorough training in all aspects of HIGH THROUGHPUT CLONING, EXPRESSION AND CRYSTALLIZATION at the Oxford Protein Production Facility (OPPP) at the start of the project, according to a standard agreement between Jonathan Grime's lab and the OPPP.
Besides, although I am experienced in structure visualisation and comparison, I have not solved the 3D structure of a protein (having worked on disordered proteins) and thus will undergo training in STRUCTURE RESOLUTION BY X-RAY CRYSTALLOGRAPHY.

- Career-enhancing skills

I will attend a series of seminars run by the Oxford Learning Institute (http://www.learning.ox.ac.uk), which are aimed specifically at the development of management and leadership skills in scientific researchers who are ready to develop an independent research program within the university.

Q17 DATA MANAGEMENT & DATA SHARING (no more than 1500 words)

Where appropriate, detail (a) your plans for data management, curation and storage; (b) your policy for sharing data with others, including the management and prioritisation of access to data; (c) your strategy for current and future communication with user communities; and (d) any ethical considerations.

(a) your plans for data management, curation and storage:
This work will generate a large curated dataset of bona fide overlapping genes (encoding proteins whose evidence has been proven experimentally). It should comprise about 200 viral and 50 bacterial overlapping genes, together with bibliographical references, annotated functions and structural features, infected host, etc.
Its curation is described in Experimental design, and involves mainly extensive bibliographical researches.
Manual curation is very time-consuming, especially since i) a significant proportion of hypothetical overlaps are spurious; ii) some bona fide overlaps are not annotated in sequence databases.
Therefore, this curated dataset will be of HIGH VALUE for biomedical researchers (see paragraph below).
However, it is well-known that most bioinformatics databases go down in a few years or are not updated (Birney and Clamp 2004, Galperin 2008). Therefore, rather than creating a database, we will make the complete dataset FREELY AVAILABLE AS A TEXT FILE.

(b) your policy for sharing data with others, including the management and prioritisation of access to data:
The dataset will be linked in the website for Robert Belshaw RNA Virus Database (http://virus.zoo.ox.ac.uk/rnavirusdb). It will also be available as a supplementary file published with the article and thus freely downloadable from the publisher web's site.

In addition, the curated dataset will serve the following goals:
- to improve several widely used sequence databases
- to develop a benchmark for overlapping gene prediction software.
- to help improve protein 3D structure prediction methods.

I will take GREAT CARE TO MAXIMISE THE UPTAKE OF THE DATASET, in the following ways:

- I will write to the curators of the NCBI viral database (www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html) to ensure that the overlapping genes that are not referenced therein are included. This will be very useful to the community since it is the most used viral sequence repository, but is not completely reliably annotated. For instance, in my previous study I found out that 10 bona fide gene overlaps were not referenced.

- Because I will carry out an evolutionary analysis on hundreds of viral proteins, I expect to discover numerous evolutionary connections between protein domains (described as "clans" in the protein family database PFAM (http://pfam.sanger.ac.uk). In my previous work, I have discovered 9 PFAM clans so expect to now discover about 30 clans.
I will thus write to the curators of PFAM to suggest inclusion of these novel clans.

- The dataset can serve as a benchmark for software designed to detect overlapping genes, which is one of Robert Belshaw's current projects. Improving such software is important; for instance the freely available overlapping gene prediction software MLOGD (http://guinevere.otago.ac.nz/aef/MLOGD/index.html) has led to the discovery of previously overlooked viral overlapping genes, including several ones essential for the viral lifecycle. I will contact the developers of such software.

- I will deposit the structures of proteins we have solved in the publicly available database PDB (www.rcsb.org). If these structures correspond to previously unknown folds, I will contact the developers of ab initio structure prediction methods to publicize our results, of great relevance to improve these methods.

(c) your strategy for current and future communication with user communities
- I will organise an INTERNATIONAL WORKSHOP ON DE NOVO PROTEIN CREATION, which will be an occasion to widely publish the dataset.
Indeed, at present some aspects of de novo protein creation are studied independently by a few teams in the world which have no mutual awareness of each other's work. With the support of my supervisors, I will invite a dozen of international researchers having each tackled different aspects of this subject to present their results, discuss, and envisage innovative collaborations. I ask for financial support for this workshop (see Miscellaneous costs), for which my experience in science communication should be invaluable.

Of course we will also publicise the existence of the dataset at congresses and to our numerous collaborators.
Should the work, at the end of this research project, finally evolve towards creation of a database of de novo proteins, we are well aware of strategies of comparable databases to communicate with user communities (for instance booths at congresses). These strategies were reviewed in an excellent special issue on databases in Briefings in Bioinformatics (Altman 2004).

Q18 OUTLINE OF PUBLIC ENGAGEMENT PLANS (no more than 250 words)

Since this is an innovative topic, of interest to, and easily understandable by the public (evolution works mainly by tinkering i.e. reusing old ideas, but also creates some novelty from scratch, just like human societies), I plan to write popular science articles and to give public talks (I worked as a public engagement professional for 7 years).

Please note that we provide support for researchers in the UK and Republic of Ireland to engage with the lay public. Do you wish to receive information about training, funding and other public engagement opportunities?

Yes

Q19 RECOMMENDATION BY APPLICANT’S PRESENT HEAD OF DEPARTMENT OR SUPERVISOR
(no more than 500 words)

Marseille, May the 13th 2009

Dr. Sonia LONGHI
Directeur de Recherches DR2, CNRS

Laboratoire dâ€™Architecture et Fonction des MacromolÃ©cules Biologiques (AFMB) -
UMR 6098, CNRS et UniversitÃ©s dâ€™Aix-Marseille I and II
163 Avenue de Luminy, Case 932, 13288 Marseille Cedex 09, France.
Tel: 33-491 82 55 80; Fax: 33-491 26 67 20;
E-mail: Sonia.Longhi@afmb.univ-mrs.fr

Letter of recommendation for Dr David Karlin's application to a
Wellcome Trust career re-entry grant

To whom it may concern

I have been Dr David Karlin's PhD supervisor and thus have closely worked
with him for more than 3 years on a project that I had just initiated when he arrived.

Although working on a newly started research project is not easy, Dr Karlin obtained results sound enough to warrant publication of five articles, with this angle of research being still pursued in the lab. Dr Karlin clearly loved science but thought there was more about it than working in a lab and publishing papers, and wanted to see how science was developing in a societal context. He thus left research to set up a science communication association, which he headed for 6 years, and which is still successful. At the time, knowing him, I jokingly told him that he would come back to research. I'm happy to see that this prediction is about to become true.

Since he left the lab to reconvert to science communication, he nevertheless has kept on demonstrating his strong interest in research by working on his spare time on a project that is a follow-up of observations he made during his PhD project, managing to author one paper that is undergoing minor revisions in J. Virol. Beyond his scientific merits, Dr Karlin's professional experience has endowed him with many other skills that should be very useful for him to resume a successful career in research, including project management, teaching, communication, networking and grants writing.

I have no doubt that his future work will be high scientific quality and will contribute to the development of the rather unexplored field of de novo proteins. For this reason I wholeheartedly endorse his application in the strongest possible term.

Please do not hesitate to contact me if I may be of further assistance.
Best regards

Sonia Longhi, PhD

Full name: Sonia Longhi Position: Principal Investigator

Role: Supervisor

Q20 RECOMMENDATION BY SUPERVISOR AT INSTITUTION WHERE AWARD WILL BE HELD
(no more than 500 words)

UNIVERSITY OF OXFORD
DEPARTMENT OF ZOOLOGY

SOUTH PARKS ROAD OXFORD OX13PS

01865 281997 - robert.belshaw@zoo.ox.ac.uk

14th May 2009

David Karlin contacted me originally because his research on overlapping genes was complementary to a study that I had published recently in Genome Research, looking at the role that overlapping genes play in genome compression. Over a series of meetings, it became clear to me that David had both exciting ideas about protein evolution, and the drive to convert these ideas into practical research projects. Our discussions have led to the research project in this Fellowship application.

David has a strong background in structural and evolutionary virology, and I am confident that his research proposal will lead to several very good publications and have a marked impact on the field. I anticipate the project also leading to other developments in the field, for example, his curated dataset of overlapping genes in more than 200 viruses, once made available on the web, will allow genome annotation software to be tested against. David's skills in communication, gained while working outside science, will be of great value in ensuring that his work is taken up by other workers. In particular, I am impressed by his plans for networking initiatives such as a workshop on de novo protein creation, and the Department here at Oxford would be a good venue for such a workshop.

The project would involve David working both in my group and that of Jonathan Grimes. This is quite feasible: David not only has experience in all aspects of the proposed research, but also and critically he has significant experience in project management. Also, I consider that his working on his research in his spare time for over four years is a sure indicator of his motivation, perseverance and good time management.

David will receive excellent training here. I will personally train him in the statistical analysis of large bioinformatic datasets. The Oxford Learning Institute (http://www.learning.ox.ac.uk) offers a comprehensive and world class range of seminars, workshops and programs to develop the careers of staff at the University of Oxford. David will attend a series of seminars run by the Institute which are aimed specifically at the development of management and leadership skills in scientific researchers who are ready to develop an independent research program within the university. He will compete for D.Phil. studentships along with other members of staff and join the department's mentoring scheme for newly appointed academics. In addition, the department has an excellent record in attracting external project funding, and David would be strongly supported by myself and other colleagues to write successful further grant applications.

I will monitor David's progress throughout the fellowship by weekly meetings, and ensure that he gives regular talks within the department on his research. He has all the qualities necessary to succeed in research and I am confident that the end of this fellowship would see him ready to set up his own research group.

Yours faithfully,

Robert Belshaw, PhD

Full name: Robert Belshaw   Position: Departmental Lecturer

Q20 RECOMMENDATION BY SUPERVISOR AT INSTITUTION WHERE AWARD WILL BE HELD

(no more than 500 words)

Dear Sir/Madame,

I am delighted to write in strong support of David Karlins application to the Wellcome Trust for a Career Re-entry Fellowship. It is clear that David is very committed to science, both his own research but also conveying of the importance of science to a modern society to the general public. Having taken a “sabbatical” from his scientific research career to become more involved in scientific public engagement, he now wishes to develop some of his scientific ideas within the structure and framework of an academic lab. His interests in the evolution of viruses and proteins mirror some of my interests, and I would be more than happy to act as a mentor to him within the Division of Structural Biology, so that he can develop his own independent scientific career. David already is very familiar with the skills required in a molecular protein structure lab, and I have no doubt he will very easily pick-up the latest tools and technologies that have been developed over the past few years, whilst he has been engaged in developing his public engagement activities.

Clearly David has a deep interest in science and in particular biology, so much so that he has been able to develop his own ideas to a point (outside of a working scientific lab) where they are now publishable. He is to be highly commended for this and I hope he is successful in pursuing a scientific career.

Full name: Jonathan Grimes Position: Research Lecturer

Q22 **INFLATION** The costs requested will be increased for inflation by the Wellcome Trust. However, the Trust would like to monitor the current inflation rate(s) that the institution would normally use when costing applications.

Salary inflation rate (% per annum):	2
Non-salary inflation rate (% per annum):	2

Q23 **SUMMARY OF FINANCIAL SUPPORT REQUESTED**

Please specify currency used	UKï¿½ Sterling
Duration of grant (state in months):	36

		TOTAL COST
(a)	Applicant's salary	142167
(b)	Materials and consumables	57000
(c)	Animals	0
(d)	Equipment	3410
(e)	Miscellaneous	24500
(f)	Work abroad	0
	GRAND TOTAL	**227077**

Q24 DETAILS OF FINANCIAL SUPPORT AND RESOURCES REQUESTED

(a) Applicant's salary

(i)	Salary grade/scale	8.1
(ii)	Basic starting salary	37628
	Total cost on grant	**142167**

(b) Materials and consumables (description)	**Costs**
1) Bioinformatics part in Robert Belshaw's lab: Computer and bureautics-related (hardware, software)	7000
2a) Experimental part in Jonathan Grimes' lab: Consumables for cloning and protein expression at Oxford Protein Purification Facility (250 expression constructs)	20900
2b) Experimental part in Jonathan Grimes' lab: Consumables for *purification*	15300
2c) Experimental part in Jonathan Grimes' lab: Consumables for *crystallisation*	13800
Subtotal	**57000**

(c) Animals	
Total purchase cost Total maintenance cost Total procedures cost Total associated cost	
Subtotal	

The table below should be duplicated for **each different species.**

Recommendations

Q24 DETAILS OF FINANCIAL SUPPORT AND RESOURCES REQUESTED (cont.)

(d) Equipment

Contact details for the Institution's Director of Procurement/Head of Purchasing (or equivalent).

Name: Mark Bowen

Tel: 018 6561 6042

Address:
Director of Purchasing
University Offices, Wellington Sq
Oxford University
Oxford
OX1 2JD

E-mail: mark.bowen@admin.ox.ac.uk

(i) Request for equipment.

Type of equipment	Equipment specification	Preferred manufacturer/ supplier (if known)	Maintenance contract duration (months)	Cost of maintenance contract	Number of items	Cost per item	**Total cost**	**Contribution from other sources**	**Amount requested**
							0		**0**
						Total:	**0**		**0**

Recommendations

Q24 DETAILS OF FINANCIAL SUPPORT AND RESOURCES REQUESTED (cont.)

(d) Equipment (cont.)

(ii) Request for equipment maintenance.

Maintenance of existing Wellcome Trust-funded equipment

The Wellcome Trust will only consider providing maintenance funds for equipment more than five years old if the applicant can demonstrate it is cost-effective to do so.

Details of equipment/facility	Wellcome Trust grant reference number of original award	Award start date	Award end date	Date of purchase	Start/end dates of any current maintenance contract & length	Total cost of Maintenance contract	% time of use for this project	**Total cost for project**
								0
							Total	**0**

(iii) Request for access charges.

Access Charges

Details of equipment/facility	Original source of funding	Wellcome Trust grant reference number, if applicable	Standard access charge per hour/day	Specify “hourly” or “daily”	Hours/days of use for this project	**Total cost for project**
Wellcome Trust Centre for Human Genetics Core facilities and IT	Core grant	075491	3.41	daily	500	**1705**
Contribution to the running costs of the Division of Structural Biology	Core grant	075491	3.41	daily	500	**1705**
						0
					Total	**3410**

Recommendations

Q24 DETAILS OF FINANCIAL SUPPORT AND RESOURCES REQUESTED (cont.)

(e) Miscellaneous (description)	**Costs**
Training for myself in high-throughput expression and crystallisation, protein bioinformatics methods, statistics applied to biomedical sciences	4500
Organisation of an international workshop on de novo creation and evolution of proteins, a new field of research (12-15 researchers, 3 days), in the UK	20000
Subtotal	**24500**

Q25 ACCESS TO RADIATION SOURCES

(a) Synchrotron Radiation Sources

(i) Will the proposed research require access to a synchrotron radiation source (SRS)? Yes

(ii) Please specify to which source(s) you will be applying

- European Synchrotron Radiation Facility (BM14 beamline)
- Diamond Light Source

(b) Neutron Sources

(i) Will the proposed research require access to a neutron source? No

(ii) Are you requesting costs from the Wellcome Trust?

(iii) If yes, complete table below (anticipated usage must be specified in whole days) and Q24 (d)(iii) Access Charges, detailing the costs required.

Details of neutron source	Total number of days	Number of days per annum				
		Year 1	Year 2	Year 3	Year 4	Year 5
	0					
	0					
	0					
	0					
	0					
	0					
	0					
Total	**0**					

(iv) Please justify your proposed access to the neutron source, including the number of days requested (no more than 500 words).

Recommendations

Q26 REASONS FOR SUPPORT REQUESTED

In this section, justify:

(a) Materials and consumables (no more than 300 words)

- computer and associated hardware and software for the bioinformatics part

- consumables for the high-throughput cloning, expression, purification and crystallization of 250 expression constructs. The figures below are derived from the mean prices charged by the Oxford Protein Purification Facility, taking into account the standard attrition rates of each step (expression, purification, crystallisation) for viral and bacterial proteins.

(b) Animals (numbers and species) (no more than 300 words)

(c) Equipment, equipment maintenance and access charges (no more than 700 words)

For access charges, please show how they have been calculated on a cost-recovery basis. This can include (i) a maintenance or service contract providing a basic level of service; (ii) running costs; (iii) materials and consumables; and (iv) staff time. Please also state the percentage of time/number of hours the equipment/facility will be used for the project.

I am requiring access charges only.

The interdisciplinary nature of the project and the use of communal facilities means that the running costs must cover not only the lab consumables but must also contribute to:

- the communal computing, experimental equipment (such as MALLS, Thermofluor, DLS ...) and specialist infrastructure of the Division of Structural Biology STRUBI (eg contribution to running costs of the wet lab infrastructure, such as centrifuges, tissue culture, in-house X-ray facilities, high performance computing, computer consumables and archive media).

- a contribution to the core facilities and IT of the Wellcome Trust centre for Human Genetics, in which STRUBI is located.

(d) Miscellaneous costs (no more than 300 words)

1) Trainings

I will undergo the trainings described in the corresponding section of the application, which are necessary due to the multidisciplinary and fast-evolving nature of the field.

2) Organisation of a workshop

At present some aspects of de novo protein creation are studied independently by a few teams in the world but they have no mutual awareness of each other's work. With the support of my supervisors, I will thus organise a 3-day workshop to help this new field of research coalesce. I will invite a dozen of international researchers having each tackled different aspects of this subject to present their results, discuss, and propose innovative collaborations.

(My previous experience in organising events such as non-scientific workshops or a science festival will be precious).

Q27 FULL ECONOMIC COSTING (UK applicants only)

The Wellcome Trust would like to monitor the full economic cost of research proposals. If your institution is calculating the full economic costs of this proposal, the table below should be completed.

Please note that the Wellcome Trust will not fund the full economic cost of research and the actual costs sought from the Wellcome Trust should be detailed in the 'DETAILS OF FINANCIAL SUPPORT AND RESOURCES REQUESTED' section of the form.

This information is being gathered for monitoring purposes only and will have no bearing on the peer review and decision-making process for your application.

(a) Does the host Institution use TRAC or an alternative methodology validated by the UK Research Councils to calculate full economic costs? Yes

(b) If yes, please complete the following table:

	Full Economic Cost (£)
Directly Incurred Costs	
Staff	142167
Travel and subsistence	0
Other costs	81500
Equipment	3410
Subtotal	**227077**
Directly Allocated Costs	
Principal Applicant salary costs	0
Coapplicant salary costs	0
Estates costs	38679
Other directly allocated costs	3768
Subtotal	**42447**
Indirect Costs	**144180**
TOTAL	**413704**

Q28 RESEARCH INVOLVING HUMAN PARTICIPANTS, BIOLOGICAL SAMPLES AND PERSONAL DATA RELATING TO LIVING OR DEAD PERSONS

(a) Does your project involve human participants? No

If yes, refer to notes.

(b) Will personal data be used? No

(c) Will your project involve use of biological samples? No

(d) Please state:

(i) By whom and when the ethics of the project has been reviewed, and specify any other regulatory approvals that have been obtained.

And/or:

(ii) By whom and when the ethics of the project will be reviewed, and specify any other regulatory approvals that will be sought.

(e) In the course of your project:

(i) Do you propose to use facilities within the National Health Service (NHS)? No

(ii) Does your research involve patients being cared for by the NHS? No

(iii) If the answer is yes to (i) or (ii) above, please indicate which organisation has agreed to be the sponsor for the project under the Research Governance Framework for Health and Social Care, published by the Department of Health in England or the corresponding departments in Northern Ireland, Scotland or Wales.

Please note that the Wellcome Trust cannot act as sponsor.

(f) If your project involves a clinical trial:

(i) Please state whether it is covered by The Medicines for Human Use (Clinical Trials) Regulations.

(ii) Please indicate which organisation has agreed to be the sponsor for the project.

Please note that the Wellcome Trust cannot act as sponsor.

Q29 EXPERIMENTS ON ANIMALS

(a) Do your proposals involve the use of animals? No

(b) Do your proposals involve the use of animal tissue? No

(c) Do your proposals include procedures to be carried out on animals in the UK which require a Home Office licence?

If yes, refer to notes.

(d) Does the institution where the animal work is to be carried out hold a certificate of designation under the Animals (Scientific Procedures) Act 1986?

(e) Do your proposals involve the use of animals or animal tissue outside the UK?

If yes, refer to notes.

(f) If your project does involve the use of animals, what would be the severity of the procedures?

(g) Please provide details of any procedures of substantial or moderate severity (no more than 250 words).

(h) Why is animal use necessary: are there any other possible approaches? (no more than 250 words)

(i) Will the following species to be used?

Primate	
Cat	
Dog	
Equidae	
Genetically Altered Animals	
Other animals	

(j) Why is the species to be used the most appropriate? (no more than 250 words)

(k) Primates

(i) Do you expect facilities and practices, and the proposed research will comply with the principles set out in the 'National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs) Guidelines: Primate accommodation, care and use' (http://www.nc3rs.org.uk/downloaddoc.asp?id=418)?

If not, please explain why.

(ii) Will it be necessary to transport the non-human primates (i.e from breeding facility and within the host institution environment)?

If so, indicate approximate journey times and the measures that will be taken to minimise the potential stress during transport.

(iii) Will single housing of the non-human primates be necessary at any time?

If so, please provide details in terms of the justification for single housing, its duration, and what additional resources will be provided to the animals to minimise the impact on animal welfare.

(iv) Describe the experimental procedures involved and how any pain, suffering, distress and/or lasting harm will be minimised. Have the procedures been recently reviewed by the Named Veterinary Surgeon (NVS), Named Animal Care and Welfare Officer (NACWO) and ethical review process (ERP)?

Administration

(v) Will any of the experimental procedures involve food and/or water restriction?

If so, justify why this is necessary and outline what alternatives have been considered.

(vi) Will any of the experimental procedures involve restraint?

What alternatives have been considered? Describe the nature of the restraint, its duration and frequency, and what will be done to avoid distress?

(vii) What prior experience and training in non-human primate use, care and welfare have the staff named in the application had? What provision is made for continuing professional development in these areas?

(viii) Will any of the staff involved require specific training for any of the procedures concerned?

Please provide details of the training needed and where it will be undertaken.

Specific questions for Cats and Dogs

(I) Cats and Dogs

(i) From where will the animals be sourced?

(ii) Will it be necessary to transport the animals?

If so, indicate approximate journey times and the measure that will be taken to minimise the potential stress during transport.

(iii) Are animals to be imported?

Where animals are to be imported, what journey times have been agreed with the Home Office? Describe the conditions for the animals at the breeding establishment and how the potential stress during transport will be minimised.

(iv) Please provide details of the housing for the animals, e.g. enclosure size, environmental enrichment.

(v) Will single housing of the animals be necessary at any time?

Administration

If so, please provide details in terms of the justification for single housing, its duration, and what additional resources will be provided to the animals to minimise the impact of the single housing.

(vi) Describe the experimental procedures involved and how any pain, suffering, distress and/or lasting harm will be minimised. Have the procedures been recently reviewed by the Named Veterinary Surgeon (NVS), Named Animal Care and Welfare Officer (NACWO) and ethical review process (ERP)?

(vii) Will any of the experimental procedures involve restraint?

What alternatives have been considered? Describe the nature of the restraint, its duration and frequency, and what will be done to avoid distress?

(viii) What prior experience and training in animal use, care and welfare will be required of the staff named in the application? What provision is made for continuing professional development in these areas?

(ivx) Will any of the staff involved require specific training for any of the procedures concerned?

Please provide details of the training needed and where it will be undertaken.

Q30 RISKS OF RESEARCH MISUSE

(a) It is the responsibility of institutions in receipt of Wellcome Trust funding to ensure that any risks that research could be misused for harmful purposes are managed in an appropriate manner.

Please confirm that you have considered whether your proposed research could generate outcomes that could be misused for harmful purposes. Yes

(b) If you have identified any tangible risks of this type, please briefly describe these risks and the steps that you and your institution will take to manage them (no more than 250 words).

Q31 LOCATION OF RESEARCH

(a) Will the research project be undertaken in a Wellcome Trust Clinical Research Facility? No

If yes, please specify:

(b) Will the research project be undertaken in the Wellcome Trust Sanger Institute or a Wellcome Trust Centre? Yes

If yes, please specify: Human Genetics, Oxford

Administration

Please provide a letter of support from the Director of the Centre/Clinical Research Facility specified.

Q32 CONSULTANCIES, EQUITIES AND DIRECTORSHIPS

Do any of the applicants have consultancies or any equity holdings in, or directorships of, companies or other organisations that might have an interest in the results of the proposed research?

No

If yes, give brief details (no more than 200 words).

Q33 COMMERCIAL EXPLOITATION

(a) Will the proposed research use technology, materials or other invention that, as far as you are aware, are subject to any patents or other form of intellectual property protection?

No

If yes, give brief details (no more than 200 words).

(b) Is the proposed research, in whole or in part, subject to any agreements with commercial, academic or other organisations?

No

If yes, give brief details (no more than 200 words).

(c) Is the proposed research likely to lead to any patentable or commercially exploitable results?

No

If yes, give brief details (no more than 200 words).

(d) If any potentially commercially exploitable results may be based upon tissues or samples derived from human participants, please confirm that there has been appropriate informed consent for such use.

Administration

ADDITIONAL INFORMATION

Additional information you wish to communicate to the Trust. For example, please state if you are sending additional material, such as collaborators' forms, under separate cover. Suggested referees should be listed in a covering letter, which can be accessed via your homepage.

I have enclosed supplemental material:

- DOC1: A figure that sums up the main findings of my article which is under minor revisions for Journal of Virology (May 2009) and that can be seen as a successful pilot for the bioinformatics part or the project.
- DOC2: a part of the unpublished curated dataset of bacterial de novo proteins that will use as "backup" to ensure success of the final phase, solving the structure of de novo proteins.
- DOC3: a part of the unpublished curated dataset of proteins suitable for X-ray crystallography, that I will use for the pilot phase at the start of the project