Proceedings of the Symposium on Synchrotron Crystallography, Krynica, Poland 2001

Structural Genomics in Europe and beyond — Shifting Scientific Directions at EMBL Hamburg

M. Wilmanns*

EMBL Hamburg Outstation c/o DESY
Notkestrasse 85, 22603 Hamburg, Germany

The emerging structural genomics initiatives provide novel opportunities to complement the rapidly increasing amount of genomic sequence data with the three-dimensional molecular structures of the coded genes. Many of these gene products exert their cellular functions by interacting with multiple partners. Unravelling the molecular structures of these interactions provides the most useful information to investigate their involvement in cellular processes. To this end, the determination of structures exceeds the number of coded gene products by several orders of magnitude. Structural genomics offers opportunities to synergise European research in structural biology technologies, with a multitude of excellent centres currently available. In this contribution, current initiatives of the Hamburg Outstation of the European Molecular Biology Laboratory will be outlined.

PACS numbers: 87.15.–v

1. Introduction

In the last few decades, the emerging capability to experimentally determine molecular three-dimensional structures of biological macromolecules at the atomic level can be seen as one of the most fascinating developments in molecular and cellular biology. Disclosure of structures, such as myoglobin and hemoglobin — the first integral membrane proteins — and some huge macromolecular complexes, like the blue tongue virus or the ribosomal subunits, have gained world-wide recognition as landmark discoveries in the life sciences. In particular, X-ray crystallography has developed to such an extent to sufficiently permit generalised application to the full range of biological macromolecules with a unique, native conformation.

*corresponding author; e-mail: Wilmanns@embl-hamburg.de

Access to an increasing network of dedicated synchrotron facilities mostly in North America, Japan, and Europe, including three third generation low emittance facilities at present, has been critical within these developments. Hence, structural biology offers significant potential to complement novel scientific directions in life sciences.

More recently, the rapidly increasing availability of genomes from different organisms has opened up unprecedented opportunities in molecular biology and medicine. Sequence data availability has shifted the investigation of biological processes from focus on single molecules, which have been frequently linked fortuitously to specific processes, to quantitative assessment of genomic (RNA, DNA) and proteomic alterations. As a spin-off, novel key technologies such as those involving quantitative immobilisation on chips or mass spectrometry determinations have emerged quickly, endowing genomic data analysis as a generally-applicable targeted research approach. Due to their technological advancement, structural biology techniques, such as X-ray crystallography and NMR spectroscopy, are rapidly becoming key tools in these new research directions, complementing available one-dimensional sequences with three-dimensional footprints. However, limited by the nature of the experiments, which impede the use of parallel multi-target approaches, these techniques presently remain focused on single molecular entities, mandating time and resource considerations. Automated structure determinations of macromolecular-small ligand complexes may be the exception to this rule provided they do not significantly affect overall shapes that determine feasibility for crystallisation. Hence, it is not surprising that these applications — screening of proteins with the known three-dimensional structure with long series of small ligands — have become the major focus of structural biology programmes within the biotech and pharmaceutical industry.

As there are no concrete expectations for novel structural biology methods, true "structural genomics" approaches can develop within the confines of massive parallelism. A task of moderate urgency concerns increasing the efficiency of established structural biology techniques by maximising their throughput at identified bottleneck stages. This challenge is primarily one of a managerial and technical nature rather than one of scientific short-comings. This approach may have a considerable impact on how science in structural biology will be conducted, diverging from a single project-single scientist approach to one exemplified by project management within large integrated teams. Therefore, it is not surprising that the issue of "structural genomics" is still somewhat controversial within the structural biology community, with many leading scientists expressing reservations as of late. The aim of this contribution is to outline where "Structural Genomics" could synergistically meet other genomics/proteomics approaches, and strengthen classical, hypothesis-driven structural biology. This will be illustrated by two projects from the Hamburg Outstation of the European Molecular Biology Laboratory (EMBL).

2. Requirements for present and future initiatives in Structural Genomics

In 1998–2000, a number of major so-called "Structural Genomics" projects, mostly in the United States but also in Germany and Japan, received considerable financial funding [1–3]. As a common denominator, emphasis has been placed on management of these "big science" projects by developing efficient infrastructures for target selection, target processing by "core facilities" aimed at high-throughput conduct at rate-limiting steps, and data dissemination by publication and deposition. In the United States, rules mandated by the National Institute of General Medical Sciences (NIGMS) have been established, by large setting reference standards for projects pursued in research facilities around the world. The primary objective of these structural genomics projects is to demonstrate the applicability of high-throughput technologies ranging from the critical steps in sample preparation (cloning, heterologous expression, purification, characterisation) to data acquisition, structure determination, refinement and interpretation [4]. Therefore, no strict requirements beyond the determinations of three-dimensional molecular structures have been imposed.

Triggered by the planned provision of resources for pilot projects in genomics/proteomics by the European Commission, intensive discussion on potential future aims of "European" Structural Genomics has already begun within the structural biology community in Europe. This exchange occurs at a time when the European Commission has announced plans to provide substantially more resources for Structural Genomics within the next 6th Framework, which is expected to start in 2003. While these discussions are still ongoing, emerging consensus on so-called "second generation" Structural Genomics projects emphasise the following needs:

- **To incorporate targets encompassing the full range of sequence and structure complexity, by including multi-functional, multi-domain targets, and targets that are integrated into biological membranes.** With the aim of developing high-throughput technologies, most early Structural Genomics projects have imposed limits on their targets — for instance by excluding targets exceeding a defined sequence length [5] or integrated membrane proteins [6] — indirectly favouring readily accessible targets known as "low hanging fruits" [7]. In contrast, more recent initiatives have focused on difficult targets, like nuclear receptor complexes [8]. In an extreme case, a new German consortium (co-ordinated by Lars-Oliver Essen, University of Marburg) will concentrate exclusively on "difficult" integral membrane protein targets. Given the complexity of handling these targets, the success of such a project will be at the expense of high-throughput applications for highly standardised methods, deviating from the objectives of most of the ongoing Structural Genomics projects. A redefinition of "Structural Genomics" may eventually be required.

- **To explicitly include protein–ligand complexes, effecting target functions in cellular processes**. Contrary to most known enzymes with a defined catalyst/substrate relation, many other protein targets may interact with multiple ligand partners, which may be spatially and temporally separated under physiological conditions. Many human diseases are caused by mutations culminating in alterations of specific protein–ligand interactions, often reflected by the accumulation of single nucleotide polymorphisms [9]. Therefore, structural identification of protein–ligand complexes rather than the protein targets only should be of paramount significance. Unfortunately, identification of these complexes is often incomplete and sometimes includes non-physiological ligands, creating a strong need to develop automated methods for *in vivo* ligand identification, with proven capabilities for false-positive tests. Indeed, seminal progress has recently been made in identifying hundreds of known and novel protein–protein complexes in yeast — considered a paradigm organism for higher eukaryotes — by employing novel techniques for the isolation of *in vivo* tagged protein–ligand complexes and mass spectrometry [10, 11]. Still, the lack-of-data on many target ligands probably defines the most severe limitation for Structural Genomics projects. Even in the limited number of cases for which physiological ligands are known, additional requirements are imposed to assess feasibility for structure solution of target–ligand complexes: (a) demonstrate that their binding is direct, rather than mediated by secondary ligands; (b) prove stoichiometric complex formation by biophysical methods; (c) the feasibility to co-crystallise target–ligand complexes.
- **To re-integrate the structural data and biological function of each target**. The significance of structural biology projects within the broad scientific framework is largely determined by their potential to reveal novel functional insights. This, however, frequently requires the application of techniques, unforeseeable prior to structural analysis. Emerging Structural Genomics projects are in a good position to provide substantial potential, as their funding is parameterised by the needs to generally apply specific methods at "genomic level" rather than by specific project requirements. Hence, the scientific directions and impact of these projects, within the wider objectives of functional genomics approaches, will be determined by strategic decisions of whether and to what extent to integrate complementary "non-structural" methods, in additional to structural core methods like X-ray crystallography or NMR-spectroscopy. Thus, Structural Genomics projects are endowed with unique opportunities to make their structural data, beyond the deposition of atomic co-ordinates, applicable towards the discoveries of novel functional insights.

- **To screen small synthetic molecule ligands for the identification of lead compounds, providing the ground for drug discovery approaches**. The majority of structural genomics initiatives are aimed at providing data to facilitate the investigation of the molecular basis of various diseases, establishing the battleground to combat them. Typically, this aim is mirrored by specific target inhibition considerations. At present, the most common approach is the identification of small molecular weight ligands, followed by their development into specific lead compounds, in some cases leading to drug discovery. In particular, NMR spectroscopy offers emerging technologies for mapping small molecule ligands onto known structures of biological macromolecules in a high-throughput mode. These methods allow the rapid disentanglement of the separate contributions of specific ligand residuals to binding, thus providing novel, site-directed opportunities for the rational drug design [12]. Therefore, a key point in the set-up of new Structural Genomics initiatives is whether and to what extent to include "non-structural" technologies, such as the provision of large compound libraries, infrastructures for combinatorial synthesis of chemical compounds, and large scale screening methods.

3. Aims for Structural Genomics by EMBL Hamburg

The current mandate of the European Molecular Biology Laboratory (EMBL), set forth in early 2001, is exemplified by a major shift in focus from research projects, which were largely driven by specific departments, towards an integrated approach aimed at playing a major role in functional genomics and proteomics approaches. Particular emphasis will be in structural biology, which always has played a strong role at EMBL and is the focus of the following four units: the programme for Structural and Computational Biology in Heidelberg; the Grenoble Outstation on the campus of the European Synchrotron Radiation Facility (ESRF); the Hamburg Outstation on the campus of the German Synchrotron Research Centre (DESY); and the European Bioinformatics Institute (or EBI Outstation) on the campus of the Saenger Centre, in Hinxton Hall near Cambridge.

Hereinafter, two examples of present and future structural genomics-oriented activities of the EMBL Hamburg Outstation will be outlined, paralleled by similar projects of the other structural biology-oriented units of EMBL. Structural genomics will supplement the following current tasks of EMBL Hamburg: (a) offer large-scale facilities in experimental structural biology by the provision of seven synchrotron radiation beam lines in protein crystallography, scattering experiments and X-ray absorption spectroscopy experiments of non-crystalline biological samples; (b) conduct research in structural biology, emphasising developments in structural biology technologies, with a current focus on software allowing automated interpretation of X-ray data [13] and scattering data [14]; and (c) provide

advance training in structural biology methods in form of two to three courses per year, complemented by Marie-Curie training fellowships funded by the European Union (http://www.emblhamburg.de/advancedtraining/mariecurie.html).

4. The giant muscle protein titin

The largest known gene product of the human genome comprises the protein titin, with a molecular weight of about 3 MDa [15]. Titin extends over one half of the sarcomere, connecting via its N-terminal region to the muscle Z-disk and by its C-terminus to the sarcomeric M-line, thereby forming the basic unit of muscle cells. It is largely associated with the two major muscle filament systems, actin and myosin, at its terminal flanking regions. Its central segment shows elastic properties, providing the molecular basis of passive elasticity of muscle sarcomeres. Because of these properties, titin is also known as "third filament" or "molecular ruler".

The majority of titin's DNA sequence was determined prior to the initiation of large-scale genome projects [16] and led to studies, which elucidated its molecular function [15]. Analysis of the complete sequence shows that, in its longest isoform it is composed of 38,183 residues from 363 exons, which predict a protein with over 300 domains [17] (Fig. 1). The majority of these domains are either folded as immunoglobulin-like (Ig) domains or fibronectin type-III domains. Because of its size, unmatched by any other gene product of the human genome, and its flexibility, there is no suitable method presently available to determine its overall three-dimensional structure at the atomic level. Therefore, our group and others have opted for an alternative approach: to dissect titin's sequence into domain segments using molecular biology techniques and then attempt to solve their structures either by X-ray crystallography or by NMR-spectroscopy. In turn, low-resolution images, currently provided by electron microscopy [18], can be used as basis to fit the high-resolution structures into the ultrastructure of titin within the muscle sarcomere. Previous studies have demonstrated how immunofluorescence techniques allow the precise localisation of these domains with respect to the ultrastructure of titin [19].

Since the first domain structure of titin — the immunoglobulin-like domain M5 from the M-band [20] — was determined within the same year its sequence was published, five other domain structures, in order from the N-terminus to the C-terminus, were solved subsequently (Fig. 1). These are as follows: the Zr7 repeat domain from the Z-disk region in complex with an EF-hand domain of actinidin-2 [21]; the Ig-like domain I1 from the I-band [22]; another Ig-like domain of the I-band I27 [23]; the structure of the FNIII domain A71 [24]; and the structure of a protein kinase domain commonly called "titin kinase" [25]. Some of these structures have had widespread implications for subsequent investigations, like, for example, the I27 Ig domain, which was used a reference for a series of studies

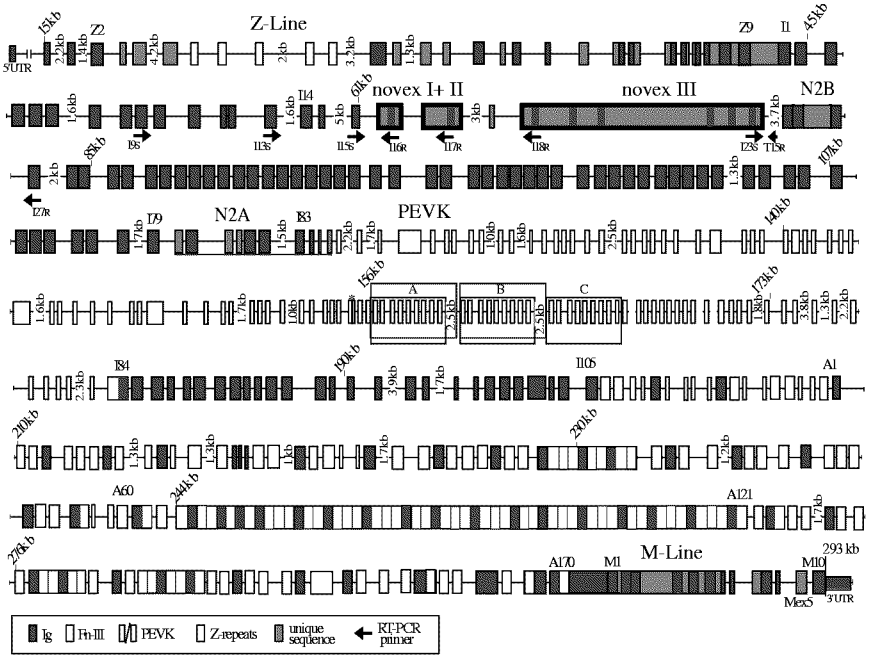

Fig. 1. Schematic domain structure of the sarcomere protein titin. The titin domains with the known three-dimensional structure are shown in red. For further information see: http://www.embl-heidelberg.de/ExternalInfo/Titin/genomic/titin-page--genomic1.html.

that examined the molecular basis of the protein's elastic properties [26]. Another example is that of titin kinase, whose three-dimensional structure allowed the identification of the molecular mechanism whereby its catalytic activity is activated by binding of calcium/calmodulin and by specific tyrosine phosphorylation [25]. In summary, high-resolution structures in the order of 2% of all titin domains are currently available.

Similar to what is done in established structural genomics projects, a substantially higher proportion of titin's domain structures can be modelled by including experimental data from analogous proteins like twitchin [27] — titin's analogue in nematodes — and by making good use of the available possibilities to model function and structure by homology [28, 29]. However, some key features, like the presence of disulphide bridges in the I1 structure [22], have not been predicted by modelling, demonstrating the need to determine many high resolution structures experimentally, even for homologous proteins.

Despite the predicted structural similarity of many domains of titin, some have been associated with domain-specific functions, as, for instance, binding to distinct ligands within the sarcomere or displaying unique biophysical properties [17]. Therefore, by treating each domain as a separate and distinct target for

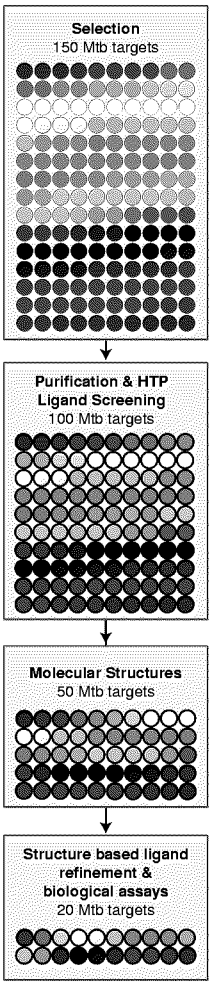

Fig. 2. Outline of the planned quantities of the Hamburg Structural Genomics consortium, illustrating the grid resolution of targets by colours. For further details, see the text.

high-resolution structure determination, titin can be regarded as small "pseudo" genome, exceeding typical virus genomes but smaller than representative bacterial genomes (in the order of 1,000–5,000 genes). All past structure determinations of titin domains have been carried out as hypothesis driven molecular projects. So far, no co-ordinated attempt has been made to fit these high-resolution data into the ultrastructure of titin and, thus, into the cellular organisation of the sarcomere. A structural genomics approach would provide an appropriate format and a systematic manner with which to bridge the gap between high-resolution domain structures and the ultrastructural and cellular organisation of titin.

5. The Hamburg TB consortium

Tuberculosis (TB) is an infectious disease of serious global threat. At present, 2 billion people world-wide are infected with the causative agent, *Mycobacterium tuberculosis* (MtB), with 8 million new cases and 2 million deaths per year. In addition, there are more 10 million people co-infected with human immunodeficiency virus (HIV), which causes AIDS, and MtB, leading to more than 0.5 million deaths per year. An increased incidence of multiple-drug resistant clinical isolates of the organism accounts for more than 3% of the TB cases in some countries. In the last couple of decades, development of new antibiotics has subsided, in part, due to the lack of sufficient economic interest by pharmaceutical companies. Recently, the complete sequencing of the MtB genome [30] has opened up novel opportunities to unravel the molecular basis of MtB pathogenicity and persistence in living organisms, promoting potential new avenues for the discovery of new specific drugs. In 1999, the EMBL at Hamburg began association with an American consortium under the co-ordination of Tom Terwilliger, at Los Alamos. The objective of the consortium was to provide structural data from about 10% of the entire MtB genome by selecting targets, which could be specifically involved in the pathogenicity of this mycobacterium.

Complementing these efforts, EMBL in conjunction with three academic partners (Max-Planck-Groups in Structural Biology, DESY, Hamburg; Max-Planck--Institute for Infectious Biology, Berlin; Institute for Bioinformatics, GSF, Neuherberg) and three industrial partners (X-ray Research, Norderstedt; Graffinity, Heidelberg; Biomax, Martinsried) from Germany has formed a structural genomics initiative with primary focus on the genome of *Mycobacterium tuberculosis*. More specifically, this project proposes to provide an integrated structural genomics approach with respect to target selection, structure determination and screening for potential lead compounds, as well as unravel the functional significance of targets in biological processes (Fig. 2).

Selection of 150 targets. The targets will be chosen by three different methods: (a) comparison between the MtB and non-pathogenic mycobacterium strains; (b) use of proteomic data to identify gene products that are up- or down-regulated under persistence conditions; and (c) use of proteomic data with infected organ material. In a subsequent step, the scientific and technical feasibility of these targets will be assessed by bioinformatics-based techniques. The selected targets will be cloned, expressed, and purified, with the aim of developing automated techniques and establishing heterologous expression in the non-pathogenic *Mycobacterium smegmatis*.

Screening for low molecular weight ligands for 100 targets. Small molecular weight compounds will first be retrieved from various databases. Following immobilisation on chips, each purified target will be screened for binding to at least ten different compounds. Four of these hits will then be used subsequently for co-crystallisation experiments and biological inhibition assays. The aim is to

solve structures from 50 targets, on average, in the presence of three different small molecular ligands. Based on these data, a second round of structure-based ligand screening will be carried out using specific in-house technologies provided by the industrial partner Graffinity.

***In vivo* characterisation of 20 targets**. Twenty targets for which the small ligand–target complex three-dimensional structures have been solved and for which the ligands have been refined by structure-based screening will be selected. Targets will undergo: (a) biological assays; (b) proteomics based analysis, testing the inhibitory effects of selected small ligands; and (c) specific MtB knock-out testing.

6. Conclusion

Structural biology technologies are providing an advanced platform for their systematic application in Structural Genomics or Proteomics Projects. The most useful information for deciphering the molecular mechanisms of cellular processes lies within the interactions of protein targets and multiple ligands that are often spatially and temporally separated. The task for present and future structural genomics projects is therefore daunting: to solve probably hundreds of thousands of physiological protein–ligand complexes, rather than a smaller number of proteins or protein domains, the feasibility of which is largely pending the development of scaleable technologies.

References

[1] T.C. Terwilliger, *Nat. Struct. Biol.* **7 Suppl.,** 935 (2000).

[2] U. Heinemann, *Nat. Struct. Biol.* **7 Suppl.,** 940 (2000).

[3] S. Yokoyama, H. Hirota, T. Kigawa, T. Yabuki, M. Shirouzu, T. Terada, Y. Ito, Y. Matsuo, Y. Kuroda, Y. Nishimura, Y. Kyogoku, K. Miki, R. Masui, S. Kuramitsu, *Nat. Struct. Biol.* **7 Suppl.,** 943 (2000).

[4] J.C. Norvell, A.Z. Machalek, *Nat. Struct. Biol.* **7 Suppl.,** 931 (2000).

[5] U. Heinemann, J. Frevert, K. Hofmann, G. Illing, C. Maurer, H. Oschkinat, W. Saenger, *Prog. Biophys. Mol. Biol.* **73,** 347 (2000).

[6] D. Christendat, A. Yee, A. Dharamsi, Y. Kluger, A. Savchenko, J.R. Cort, V. Booth, C.D. Mackereth, V. Saridakis, I. Ekiel, G. Kozlov, K.L. Maxwell, N. Wu, L.P. McIntosh, K. Gehring, M.A. Kennedy, A.R. Davidson, E.F. Pai, M. Gerstein, A.M. Edwards, C.H. Arrowsmith, *Nat. Struct. Biol.* **7,** 903 (2000).

[7] A.M. Edwards, C.H. Arrowsmith, D. Christendat, A. Dharamsi, J.D. Friesen, J.F. Greenblatt, M. Vedadi, *Nat. Struct. Biol.* **7 Suppl.,** 970 (2000).

[8] J.P. Renaud, D. Moras, *Cell. Mol. Life Sci.* **57,** 1748 (2000).

[9] S. Sunyaev, W. Lathe III, P. Bork, *Curr. Opin. Struct. Biol.* **11,** 125 (2001).

[10] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sorensen, J. Matthiesen, R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W. Hogue, D. Figeys, M. Tyers, *Nature* **415,** 180 (2002).

[11] A.C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga, *Nature* **415,** 141 (2002).

[12] P.J. Hajduk, R.P. Meadows, S.W. Fesik, *Q. Rev. Biophys.* **32,** 211 (1999).

[13] A. Perrakis, R. Morris, V.S. Lamzin, *Nat. Struct. Biol.* **6,** 458 (1999).

[14] D.I. Svergun, M.V. Petoukhov, M.H. Koch, *Biophys. J.* **80,** 2946 (2001).

[15] J. Trinick, L. Tskhovrebova, *Trends. Cell. Biol.* **9,** 377 (1999).

[16] S. Labeit, B. Kolmerer, *Science* **270,** 293 (2095).

[17] M.L. Bang, T. Centner, F. Fornoff, A.J. Geach, M. Gotthardt, M. McNabb, C.C. Witt, D. Labeit, C.C. Gregorio, H. Granzier, S. Labeit, *Circ. Res.* **89,** 1065 (2001).

[18] J.M. Squire, *Curr. Opin. Struct. Biol.* **7,** 247 (1997).

[19] W.A. Linke, *Adv. Exp. Med. Biol.* **481,** 179 (2000).

[20] M. Pfuhl, A. Pastore, *Structure* **3,** 391 (1995).

[21] R.A. Atkinson, C. Joseph, G. Kelly, F.W. Muskett, T.A. Frenkiel, D. Nietlispach, A. Pastore, *Nat. Struct. Biol.* **8,** 853 (2001).

[22] O. Mayans, J. Wuerges, S. Canela, M. Gautel, M. Wilmanns, *Structure* **9,** 331 (2001).

[23] S. Improta, A.S. Politou, A. Pastore, *Structure* **4,** 323 (1996).

[24] C. Muhle-Goll, M. Nilges, A. Pastore, *J. Biomol. NMR* **9,** 2 (1997).

[25] O. Mayans, P.F. van der Ven, M. Wilm, A. Mues, P. Young, D.O. Furst, M. Wilmanns, M. Gautel, *Nature* **395,** 863 (1998).

[26] P.E. Marszalek, H. Lu, H. Li, M. Carrion-Vazquez, A.F. Oberhauser, K. Schulten, J.M. Fernandez, *Nature* **402,** 100 (1999).

[27] G.M. Benian, X. Tang, T.L. Tinley, *Adv. Biophys.* **33,** 183 (1996).

[28] F. Fraternali, A. Pastore, *J. Mol. Biol.* **290,** 581 (1999).

[29] C.C. Witt, N. Olivieri, T. Centner, B. Kolmerer, S. Millevoi, J. Morell, D. Labeit, S. Labeit, H. Jockusch, A. Pastore, *J. Struct. Biol.* **122,** 206 (1998).

[30] S.T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M.A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J.E. Sulson, K. Taylor, S. Whitehead, B.G. Barrell, *Nature* **393,** 537 (1998).