Nepogodiev, J. F. Stoddart, D. J. Williams, *Angew. Chem.* **1997**, *109*, 1617; *Angew. Chem. Int. Ed. Engl.* **1997**, *36*, 1451.

[6] S. Höger, K. Bonrad, A. Mourran, U. Beginn, M. Möller, *J. Am. Chem. Soc.* **2001**, *123*, 5651.

[7] a) J. Zhang, J. S. Moore, *J. Am. Chem. Soc.* **1992**, *114*, 9701; b) A. S. Shetty, J. Zhang, J. S. Moore, *J. Am. Chem. Soc.* **1996**, *118*, 1019; c) Y. Tobe, N. Utsumi, K. Kawabata, K. Naemura, *Tetrahedron Lett.* **1996**, *37*, 9325; d) Y. Tobe, N. Utsumi, A. Nagano, K. Naemura, *Angew. Chem.* **1998**, *110*, 1347; *Angew. Chem. Int. Ed.* **1998**, *37*, 1285; e) S. Tanikawa, J. S. Moore, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **1999**, *40*(2), 169; f) S. Lahiri, J. L. Thompson, J. S. Moore, *J. Am. Chem. Soc.* **2000**, *122*, 11315.

[8] a) S. Höger, *J. Polym. Sci. Part A: Polym. Chem.* **1999**, *37*, 2685; b) M. M. Haley, J. J. Pak, S. C. Brand, *Top. Curr. Chem.* **1999**, *201*, 81; c) P. Siemsen, R. C. Livingston, F. Diederich, *Angew. Chem.* **2000**, *112*, 2740; *Angew. Chem. Int. Ed.* **2000**, *39*, 2632.

[9] J. C. M. van Hest, D. A. P. Delnoye, M. W. P. L. Baars, C. Elissen-Román, M. H. P. van Genderen, E. W. Meijer, *Chem. Eur. J.* **1996**, *12*, 1616.

[10] J. S. Moore, S. I. Stupp, *Macromolecules* **1990**, *23*, 65.

[11] a) H. Kukula, U. Ziener, M. Schöps, A. Godt, *Macromolecules* **1998**, *31*, 5160; b) M. Lee, B.-K. Cho, H. Kim, W.-C. Zin, *Angew. Chem.* **1998**, *110*, 661; *Angew. Chem. Int. Ed.* **1998**, *37*, 638; c) L. H. Radzilowski, S. I. Stupp, *Macromolecules* **1994**, *27*, 7747; d) D. Marsitzky, T. Brand, Y. Geerts, M. Klapper, K. Müllen, *Macromol. Rapid Commun.* **1998**, *19*, 385; e) M. A. Hempenius, B. M. W. Langeveld-Voss, J. A. E. H. van Haare, R. A. J. Janssen, S. S. Sheiko, J. P. Spatz, M. Möller, E. W. Meijer, *J. Am. Chem. Soc.* **1998**, *120*, 2798.

[12] **1a** forms under the same conditions a suspension.

[13] The systems presented here do not have thermotropic liquid crystalline characteristics. However, polysiloxane-substituted rings show this behavior: S. Höger, S. Rosselli, unpublished results.

[14] a) H. Hoffmann, G. Hebert, *Angew. Chem.* **1988**, *100*, 933; *Angew. Chem. Int. Ed. Engl.* **1988**, *27*, 902; b) D. J. Abdallah, R. G. Weiss, *Adv. Mater.* **2000**, *12*, 1237.

[15] CONTIN is a general purpose constrained-regularization program for inverting noisy linear algebraic and integral equations: S. W. Provencher, *Comput. Phys. Commun.* **1982**, *27*, 229; S. W. Provencher, *Makromol. Chem.* **1979**, *180*, 201.

[16] The concentrations of the samples analyzed were 0.05, 0.093, and 0.113 wt. %.

[17] The persistence length l is a measure for the stiffness of the polymer, and is defined as $\langle \cos\theta(s) \rangle = \exp(-s/l)$, with contour length s. In the case of a supramolecular aggregate, the calculation using the formula above leads to the "virtual" persistence length. For comparison, the persistence length of simple synthetic polymers is less than 1 nm, whereas the persistence length of DNA is in the range of 100 nm; A. Y. Grosberg, A. R. Khokhlov, *Giant Molecules*, Academic Press, London, **1997**.

[18] For the estimation of the rod length L from R_h, the relation $1/\sqrt{3}(\ln(L/d) + 0.38)$ was used (valid for a rodlike object of length L and diameter d). In our case we assume that $d \ll L$ because of the high value of R_h. If the ratio L/d is about 40, the ratio $R_g/R_h = 2.34$ can be used. With $R_g^2 = L^2/12$ follows $L \approx [12(2R_h)^2]^{1/2}$; a) M. Schmidt, W. H. Stockmayer, *Macromolecules* **1984**, *17*, 509; b) H.-G. Elias, *Makromoleküle, Bd. 1*, Hüthig & Wepf, Basel, **1990**.

[19] All form factors of cylindrical objects without polydispersity in the diameter exhibit pronounced oscillations in the investigated q range.

[20] The fit function is given by the expression (1).

$$\int_0^{\frac{\pi}{2}} \int_0^{\frac{R_a+R_i}{2}} \int_{\frac{R_a+R_i}{2}}^{\infty} [\{4r_a j_1(qr_a\sin(\alpha)) - 4r_i j_1(qr_i\sin(\alpha))\}\sin(ql\cos(\alpha/2))/$$
$$\{q\sin(\alpha)ql\cos(\alpha)(r_a^2 - r_i^2)\}]^2 \sin(\alpha) \qquad (1)$$
$$\exp[-(r_a - R_a)^2/(2\Delta r_a^2) - (r_i - R_i)^2/(2\Delta r_i^2)] \, dr_a \, dr_i \, d\alpha$$

A better fit is obtained by a more sophisticated electron density profile in the range $2 < d < 4$ nm. However, a precise analysis of the electron density requires more accurate scattering data.

[21] Preliminary investigations show that solid samples obtained from solution by very slow solvent evaporation ("equilibrium conditions")

[22] At the same time some of the block copolymer forms an undulating background structure on the mica.

[23] a) S. Höger, V. Enkelmann, *Angew. Chem.* **1995**, *107*, 2917; *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 2713; b) S. Höger, V. Enkelmann, K. Bonrad, C. Tschierske, *Angew. Chem.* **2000**, *112*, 2355; *Angew. Chem. Int. Ed.* **2000**, *39*, 2268.

[24] a) P. Dziezok, S. S. Sheiko, K. Fischer, M. Schmidt, M. Möller, *Angew. Chem.* **1997**, *109*, 2894; *Angew. Chem. Int. Ed. Engl* **1997**, *36*, 2812; b) A. D. Schlüter, J. P. Rabe, *Angew. Chem.* **2000**, *112*, 860; *Angew. Chem. Int. Ed.* **2000**, *39*, 864.

form different superstructures, indicating that the solution studies cannot be transferred to the morphology in the solid state.

From Structure to Function: A New Approach to Detect Functional Similarity among Proteins Independent from Sequence and Fold Homology

Stefan Schmitt, Manfred Hendlich, and Gerhard Klebe*

The explosion in information regarding protein sequence and structure demands new computational tools for the direct inference of protein function. The molecular function of proteins is almost invariably linked with the specific recognition of substrates and endogenous ligands in given binding pockets; proteins of related function should therefore share comparable recognition pockets. We have developed a new approach that is based on the placement of physicochemical descriptors assigned to the exposed binding-site residues. The deduced property descriptors can be used to retrieve common sub-structures and, thereby, related binding pockets. The solutions obtained are scored by comparing similarly exposed surface patches assigned to the same property, thus allowing detection of functional relationships among proteins independent of a particular sequence or fold homology.

The sequencing of the human genome represents only the first step toward understanding the functional and structural interplay of proteins in biological systems. Through methodological developments in "structural genomics", such as high-throughput X-ray crystallography,[1] we will be increasingly confronted with 3D structures of proteins whose actual biochemical function has yet to be assigned. In addition, structure prediction and homology modeling techniques might mature to a state where the overall protein geometry can be predicted correctly from sequence.[2] Thus, methods to infer protein function from 3D structure are desperately required. Since protein function is not necessarily confined to a particular fold or vice versa,[3] this assignment is by no means straightforward.

[*] Prof. Dr. G. Klebe, Dr. S. Schmitt, Dr. M. Hendlich
Philipps-Universität Marburg
Institut für Pharmazeutische Chemie
Marbacher Weg 6, 35032 Marburg (Germany)
Fax: (+49) 6421-282-8994
E-mail: klebe@mailer.uni-marburg.de

Protein function is often intimately connected with the recognition and/or chemical modification of endogenous ligands or substrates.[4, 5] Recognition usually occurs in well-characterized clefts or cavities of the protein surfaces.[6] In enzymes, for example, elementary steps along chemical pathways require a strictly defined spatial arrangement of the reaction partners. This, in turn, demands a corresponding highly conserved spatial arrangement of molecular recognition determinants in the protein to accommodate and spatially arrest them. The idea that a molecular recognition pattern may be conserved between binding sites of proteins of similar function motivated us to develop a new approach to comparing protein structures. Rather than considering complete proteins in terms of sequences, we search for similarities of surface-exposed physicochemical properties within binding pockets.

Several methods have been described to locate depressions on protein surfaces as putative binding sites.[7–11] We used the program LIGSITE[12] and implemented its algorithm into the protein–ligand database RELIBASE[13, 14] to detect and retrieve binding cavities from the entire body of resolved crystal structures. In the analysis, each protein considered is embedded in a regularly spaced grid. Any lattice intersection coinciding with a protein atom, or within its van der Waals radius, is discarded and the remaining lattice points are scored according to their degree of burial in the surface depressions. Adjacent lattice points that are deeply buried are clustered together to reveal contiguous cavities, which are deposited in the new object-oriented database CAVBASE. For each cavity the surface is approximated by assigning the surface-contacting intersections of the initially embedded lattice that was generated in the cavity-extraction step. The atomic coordinates of the amino acids flanking the cavity are reduced to a set of generic pseudocenters, which are classified according to five properties essential for molecular recognition: hydrogen-bond donor, hydrogen-bond acceptor, mixed donor/acceptor (for example, hydroxyl groups or side-chain nitrogen atoms on histidine), hydrophobic aliphatic, and aromatic contacting group.[15] These pseudocenters are further examined for their surface exposure and assigned to the nearest lattice surface

intersections (Figure 1). The cavity shape, the set of assigned descriptors of exposed recognition properties, and the corresponding surface patches are stored in CAVBASE together with information on the individual cavity occupants. The data stored are fully integrated with RELIBASE by using common pointers for the protein and ligand features (for example, atoms, residues) to interlink with corresponding objects in RELIBASE. A sample of 31 441 cavities was retrieved from 8308 protein structures. The new database is equipped for interactive visualization of cavity information and of results from a cavity comparison.

The stored information allows for fast and efficient comparisons within large data sets. A clique algorithm is applied to detect common sub-graphs generated by nodes corresponding to pairs of pseudocenters of equivalent properties and similar mutual distances.[16] Appropriate tolerances have been incorporated to consider structural variations resulting from conformational flexibility of the protein (for example, ligand-induced fit, domain mobility) and inherent limitations of the accuracy of determining protein structures. The multiple solutions generated are ranked by scoring their corresponding matches in terms of the assigned surface-exposed physicochemical properties.

The direct comparison of two mean-sized cavities (ca. 800 Å^3) requires about 100 s CPU time on a state-of-the-art single processor.[17] The approach is easy to parallelize. Thus, our method is capable of scanning a cavity of interest ("query cavity") against a sample of several thousand binding pockets. To reduce the original sample set in CAVBASE to a manageable size of 5000–6000 entries we filtered the data by criteria such as average cavity size and resolution of the input proteins. In the searches described below, we further discarded any cavities of proteins sharing high sequence identity with the query cavity protein to exclude trivial solutions. As a result, the top-ranked solutions exhibit both, local-surface similarity and shared binding motifs with the query cavity.

As this approach is entirely based on physicochemical properties exposed at the surface of a binding pocket it allows detection of relationships independent of a particular sequence or fold homology; it is entirely based on physico-
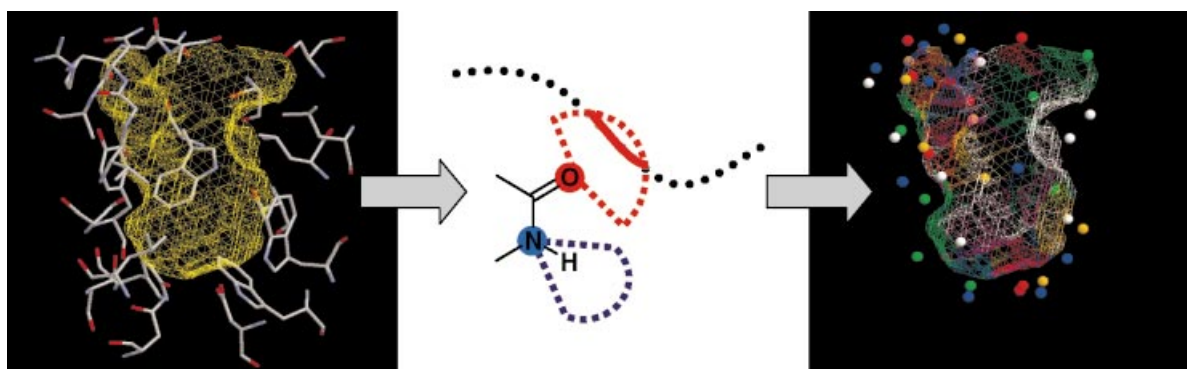

Figure 1. An extracted binding-site cavity is represented by a set of well-placed pseudocenters describing the physicochemical properties of the exposed binding-site residues flanking the cavity (H-bond donor (blue), H-bond acceptor (red), ambivalent donor/acceptor (green), hydrophobic aliphatic (white), and aromatic (orange) property; left). Only those centers that expose their property towards the cavity surface are considered (shown for the acceptor and donor properties of a peptide binding; here the donor group cannot produce a pseudocenter, because it is not exposed to the surface; center). The individual surface patches that describe the accessibility of the pseudocenters are stored and subsequently used to score the similarity among binding pockets. According to this protocol, the entire cavity is transformed into a set of property-labeled pseudocenters and associated surface patches (same color coding for pseudocenters (spheres) and the surface patches (net); right).
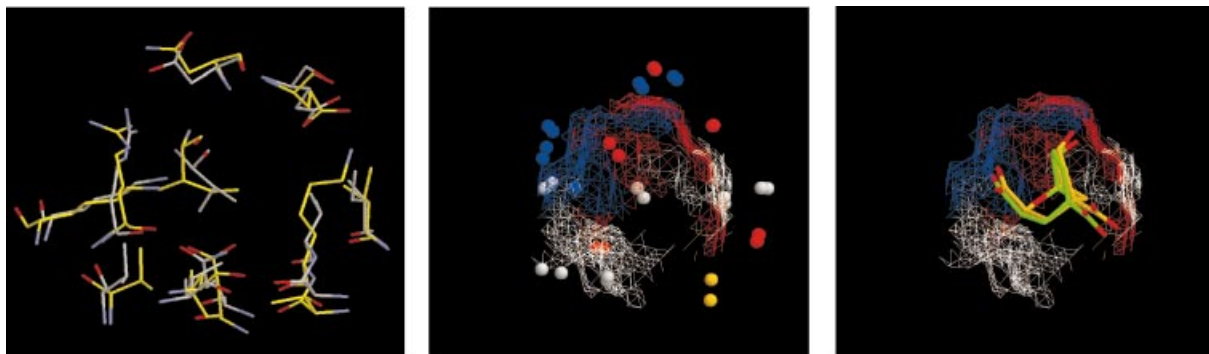
Figure 2. Binding pockets of chorismate mutase from *S. cerevisiae* (4csm) and *E. coli* (1ecm) showing less than 20 % sequence homology. Bound ligands are recognized mostly through interactions experienced by side-chain residues (left). The pocket from *E. coli* was screened against a set of 5445 different pockets using our asscoiated pseudocenter description (center). On the right the superimposed cavitites are shown together with the matched common surface patches and the bound bicyclic transition-state-analogue inhibitors.

chemical properties rather than amino acid complementarity. This is in contrast to other methods described[18–23] that are dependent on information merely retrieved directly or indirectly from the protein fold. Recently, Stahl et al.[24] reported the analysis of a preselected set of 176 zinc-containing metalloproteinases that had been clustered using a self-organizing neural network by means of solvent-accessible surface patches assigned to physicochemical properties. The active sites could be discriminated from other depresssions on the surfaces of these enzymes with their approach.

In our first example we compared the binding pockets extracted from two chorismate mutases originating from the species *Saccharomyces cerevisiae* and *Escherichia coli*.[25, 26] This example was also studied by Rosen et al.[19] by using sparse critical points[27] derived from the Connolly surface of manually extracted binding pockets. Although they show sequence identity below 20 %, both adopt a similar fold. The

bound ligand, a bicyclic transition-state-analogue inhibitor, is recognized mostly through side-chain interactions. A sample set of 5445 cavities containing the chorismate mutase cavity from *E. coli* was screened, but excluding that from *S. cerevisiae*, which instead was used as the query cavity. The cavity from the *E. coli* structure was found as the best-scored solution; Figure 2 shows the matching pseudocenters and corresponding surface patches. Although no information on the bound ligand was used, the surface match generates a transformation that results in a virtually perfect superposition of the bound inhibitors.

In a second example we probed the binding pocket from trypsin against a set of 5248 cavities taken from proteins with sequence identities below 35 % of the parent structure. The best cavities came from other members of the trypsin family, which exhibited steadily decreasing homology with trypsin. At rank 114, thus among the first 3 % of the data sample
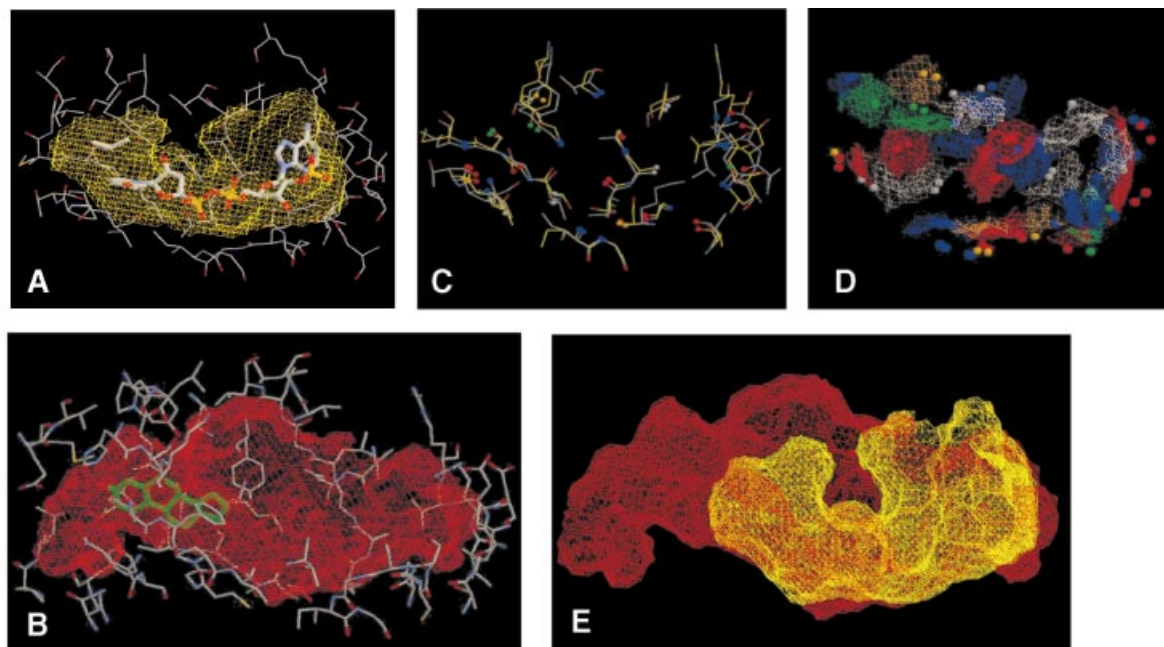

Figure 3. The NADPH cofactor binding pocket of a carbonyl reductase (1cyd; A) has been screened against a set of 5377 cavities (C and D illustrate the equivalent amino acids, pseudocenters, and surface patches, respectively). The algorithm detects a similar pocket present in a NADPH-dependent steroid dehydrogenase at rank 41 (1fds; B). The crystal structure of the steroid dehydrogenase has been determined without the bound cofactor; however, the cofactor cavity observed in the carbonyl reductase (A, yellow) matches well with a large unoccupied pocket in the steroid dehydrogenase (B, red) overlay (E).

considered, a binding pocket from the structurally unrelated subtilisin family was detected. Further examples from this family followed together with other members of the trypsin family. The trypsin and subtilisin superfamily are representatives of the major class of serine proteases, that is, they share functional similarity without sequence and fold homology.

The entire NADPH cofactor binding pocket of a carbonyl reductase[28] was taken as the query and screened against a set of 5377 cavities extracted from proteins with no significant sequence homology. The top-scoring cavities also accommodate NADPH; subsequent cavities host ligands with decreasing similarity, but which contain parts of the NADPH skeleton. The large cavity of an NADPH-dependent steroid dehydrogenase[29] is detected at rank 41 (Figure 3), although the crystal structure had been determined in the absence of the bound co-factor. A cavity from a phenol hydrolase is found at rank 116 which accommodates FAD as a co-factor. In this case, the adenine recognition site is shared with the original NADPH query pocket.

Surprisingly, in another search we detected a surface patch of an adenine-binding pocket that was similar to an unoccupied binding-site region in HIV protease. In the proteinase structure, a macrocyclic peptidomimetic is bound to the active site,[30] and leaves a binding-site region that is similar to a patch in the catalytic subunit of protein kinase A unoccupied.[31] In the latter case, this patch accommodates the adenine portion of adenylimino diphosphate. This surprising finding is of potential interest for drug design, as an adenyl moiety might be used to supplement the macrocycle in the unoccupied binding niche in HIV protease. It is likely that a large database of binding-site cavities can be used to generate interesting suggestions for new molecular portions that can be used as potential bioisosters in skeletons of novel leads.

Received: March 19, 2001 [Z16803]

[1] V. S. Lamzin, A. Perrakis, *Nat. Struct. Biol.* **2000**, *7*, 978–981.
[2] S. Roberto, U. Pieper, F. Melo, N. Eswar, M. A. Marti-Renom, M. S. Madhusudh, *Nat. Struct. Biol.* **2000**, *7*, 986–990.
[3] J. M. Thornton, A. E. Tod, D. Milburn, N. Borkakoti, C. A. Orengo, *Nat. Struct. Biol.* **2000**, *7*, 991–994.
[4] S. L. Moodie, J. B. Mitchell, J. M. Thornton, *J. Mol. Biol.* **1996**, *263*, 486–500.
[5] R. E. Babine, S. L. Bender, *Chem. Rev.* **1997**, *97*, 1359–1472.
[6] F. K. Pettit, J. U. Bowie, *J. Mol. Biol.* **1999**, *285*, 1377–1382.
[7] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, T. E. Ferrin, *J. Mol. Biol.* **1982**, *161*, 269–288.
[8] R. A. Laskowski, *J. Mol. Graph.* **1995**, *13*, 2735–2748.
[9] G. L. Levitt, L. J. Banaszak, *J. Mol. Graph.* **1992**, *10*, 229–234.
[10] J. Liang, H. Edelsbrunner, C. Woodward, *Protein Sci.* **1998**, *7*, 1884–1897.
[11] K. P. Peters, J. Fauck, C. Frommel, *J. Mol. Biol.* **1996**, *256*, 201–213.
[12] M. Hendlich, F. Rippmann, G. Barnickel, *J. Mol. Graph.* **1997**, *15*, 359–363.
[13] K. Hemm, K. Aberer, M. Hendlich, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1995**, *3*, 170–178.
[14] M. Hendlich, *Acta Crystallogr.* **1998**, *54*, 1178–1182.
[15] The generic pseudocenters are placed in a way to best represent the interaction properties of a particular functional group exposed to the cavity surface. For example, a H-bond acceptor pseudocenter assigned to a carbonyl group is given by the coordinates of the carbonyl oxygen atom, and the aromatic properties of a phenylalanine residue by the geometric mean of the six phenyl carbon atoms. Furthermore, such a center is only considered if it could perform an interaction to a

[16] C. Bron, J. Kerbosch, *Commun. ACM* **1973**, *16*, 575–577.
[17] A Pentium III processor (650 MHz) and a Linux operating system was used; CPU = central processing unit.
[18] D. Fischer, H. Wolfson, S. L. Lin, R. Nussinov, *Protein Sci.* **1994**, *3*, 769–778.
[19] M. Rosen, S. L. Liang, H. Wolfson, R. Nussinov, *J. Mol. Biol.* **1998**, *11*, 263–277.
[20] D. Fischer, R. Norel, H. Wolfson, R. Nussinov, *Proteins* **1993**, *16*, 278–292.
[21] A. C. Wallace, N. Borkakoti, J. M. Thornton, *Protein Sci.* **1997**, *6*, 2308–2323.
[22] P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, D. W. Rice, P. Willett, *J. Mol. Biol.* **1994**, *243*, 327–344.
[23] R. B. Russell, *J. Mol. Biol.* **1998**, *279*, 1211–1227.
[24] M. Stahl, C. Taroni, G. Schneider, *Protein Eng.* **2000**, *13*, 83–88.
[25] A. Y. Lee, P. A. Karplus, B. Ganem, J. Clardy, *J. Am. Chem. Soc.* **1995**, *117*, 3627–3628.
[26] N. Strater, G. Schnappauf, G. Braus, W. N. Lipscomb, *Structure* **1997**, *5*, 1437–1452.
[27] S. L. Lin, R. Nussinov, D. Fischer, H. J. Wolfson, *Proteins* **1994**, *18*, 94–101.
[28] N. Tanaka, T. Nonaka, M. Nakanishi, Y. Deyashiki, A. Hara, Y. Mitsui, *Structure* **1996**, *4*, 33–45.
[29] R. Breton, D. Housset, C. Mazza, J. C. Fontecilla-Camps, *Structure* **1996**, *4*, 905–915.
[30] J. L. Martin, J. Begun, A. Schindeler, W. A. Wickramasinghe, D. Alewood, P. F. Alewood, D. A. Bergman, R. I. Brinkworth, G. Abbenante, D. March, R. C. Reid, D. Fairlie, *Biochemistry* **1999**, *38*, 7978–7988.
[31] D. Bossemeyer, R. A. Engh, V. Kinzel, H. Ponstingl, R. Huber, *EMBO J.* **1993**, *12*, 849–859.

Unusual Formation of an Azaphospholene from 1,3,4,5-Tetramethylimidazol-2-ylidene and Di(isopropyl)aminophosphaalkyne**

F. Ekkehardt Hahn,* Duc Le Van, Michelle C. Moyes, Thorsten von Fehren, Roland Fröhlich, and Ernst-Ulrich Würthwein

During recent studies of the reactivity of N-heterocyclic carbenes[1] towards phosphaalkynes, we found that the anellated compound *N,N′*-bis(2,2-dimethylpropyl)benzimid-

[*] Prof. Dr. F. E. Hahn, Dr. D. Le Van, M. C. Moyes, T. von Fehren
Anorganisch-Chemisches Institut der Universität
Wilhelm-Klemm-Strasse 8
48149 Münster (Germany)
Fax: (+49) 251-833-3108
E-mail: fehahn@uni-muenster.de

Dr. R. Fröhlich, Prof. Dr. E.-U. Würthwein
Organisch-Chemisches Institut der Universität
Corrensstrasse 40, 48149 Münster (Germany)