On a surprising bridge between morphometrics and bioinformatics

Fred L. Bookstein

University of Vienna, University of Michigan

Abstract

Molecules do not only have formulas and enter into reactions; they also have arrangements within the body. The current tools of morphometrics are not adequate to specify those arrangements. The present essay presents a modest proposal for beginning to improve the situation: augmenting deformation analysis by a simple exterior calculus.

1 Introduction

Bioinformatics begins with anatomy, and so does this short essay. Work on human anatomy was revitalized by the NLM's Visible Human program in the 1990's. Two canonical specimens, "Adam" and "Eve," have been painstakingly digitized and most named structures segmented down to the scale of actual voxels (300x300x1000 μm, for Adam; 300 $\mu\mathrm{m}^3$, for Eve). These specimens are available in many fine software contexts for pedagogical and comparative use. The Visible Humans, in turn, energized a separate effort out of Cornelius Rosse's group at Seattle to turn the classic *Terminologia Anatomica* (the Latinate data base of anatomical labels) into a formal ontology by which concepts can be delineated and their implications and connotations tracked automatically: the Foundational Model of Anatomy [FMA]. In a typical interaction, a term from the FMA is linked to a segmented object within the image space of Adam or Eve, so that the visible surface can be queried for medical implications, or perhaps the relations of a specific anatomical structure can be visualized as if in an atlas.

These resources correspond to a relatively abstract geometrical context, in which entities have either no geometry at all (the FMA, sensu stricto) or else only appear in one or a very few instantiations not further compared among themselves (the Visible Human series). To support the principal professional context of anatomy, which was and remains the detection and classification of human variability (for diagnosis, for surgery, for cross-species comparison), such abstraction needs to be replaced by a grasp of shapes, sizes, and spatial relationships in the real organism. Yet issues like these have almost wholly fallen off the screen of current biometrical concerns. If the knowledge base of normal variability is not to die with the current generation of anatomy professors emeritus, it will have to be web-enabled, which means that it will have to be assembled by semiautomated search engines as they scan the entire corpus of extant medical literature. That corpus of literature, and the associated informatic problems, provide the other springboard for this essay.

This web-enablement of anatomy should be seen as a mandatory technology for building the bridge between future biomedical research and future medical practice. From the diagnosis of an "abnormal" ventricle to a "suspicious" pattern in an endoscopy, most decisions in the course of clinical imaging involve implicit matches between the clinical scene and a prior data base of the "normal range." Also, any bioinformatic assertion (a gradient of molecular densities, the conditions of activation of pathways) ought to be localized, sometimes to very thin sheets or tiny regions. The histological tissue type is the locus classicus of these regionalizations, but tissue types are keyed to landmarks only rarely.

2 After landmarks: coordinate systems

This is odd, inasmuch as the biomathematical construction that is the landmark was explicitly intended for the quantitative comparison of anatomies. A landmark point, for instance, is a Cartesian point (i.e. three coordinates in space in some coordinate system) along with a claim of biological correspondence to a point with the same name in every other form of a data set. In this formalism—a landmark as a Cartesian duple or triple in a biomathematical model—the coordinate system per se does not have *any* meaning; it is only the landmark configuration that does. The method of Procrustes shape coordinates that turns these configurations into a vector of variable values is well-known to this LASR audience.

These analyses are not limited to discrete geometric points (though those remain the most frequently encountered application). The algebra of a deformation engine (the thin-plate spline) conveniently associated with landmark points has been combined with multivariate statistics in a unified methodology extending the idea of biomathematically meaningful pointwise correspondence to smooth curves and smooth surfaces. A curve can be represented as a series of *semilandmarks,* landmark-like points free to slide along it, that could plausibly be claimed to correspond as points across a sample of forms; likewise a surface can be represented as a mesh of semilandmarks sliding in two directions. The curves and surfaces can be pooled as data as long as the obvious ordering relationships are enforced (e.g., surface semilandmarks can't slide across curves).

Yet even this extension of the language of landmarks is insufficient for the recording of anatomical arrangement, in view of its continued failure to span the formalizations of coordination actually required (Bookstein, 2004). Sometimes, as in biomechanics, the laboratory reference frame may matter, and sometimes, as in surgery, parts are missing or abnormally intertwined. But even in ordinary comparative contexts the ways in which information is verbalized for the current medical literature extend to derivatives as well. There's information about orientation (of photomicrographs), subdivision (medial or distal part of X), scale (the scale bar is right there, after all), clarity of boundaries (if the image is segmented automatically), and lots of similar associated quantitative channels. Much besides the stereotyped Latin is actually informative about the findings we want to turn into symbolic assertions. These ancillary anatomic terms are informal, quite distant from the raw models by which they are navigated, and not at all standardized.

The verbal descriptors that apply to these extended coordinate domains are words that specify relative locations and directions or derivatives—generalizations of anterior/posterior/medial/lateral and their equivalent when modified by directions of local features (i.e. "along the vessel"). The biomathematics of these relatively simple schemes is under control, at least as far as the differentiable models are concerned, with applications so far mainly to the human and primate brain. The phrase "under control" is meant both computationally and statistically. Generally speaking, the representation of static anatomy works either as a dependent variable (group A has different size/shape/... than group B) or as a covariate (adjusting for shape variation, group A has different average image content in the region delimited by the curve X in standard coordinates). In support of this second goal, for instance, the computer package `Edgewarp` warps exterior derivatives approximately covariantly in the course of navigations.

By contrast, there is hardly any work to build on corresponding to the computational and semantic aspects of raw anatomical variability for the nondifferentiable part of this, the specification of the "arrangement of parts". We are pretty much at the limits of all these techniques for the specification of geometry in the course of applications papers. The detailed maps involved in distributed bioinformatic computing, at whatever level of spatial scale, go well beyond existing landmark formalisms; new tools are needed to bridge morphometrics and bioinformatics in

new ways.

Even this moderate articulation of a (partial) ontology for static structures does not extend to the case of dynamic (time-dependent) anatomy that suits the even richer bioinformatic information base of pathways and organized developmental sequences of molecular states. The obvious language of change in derived quantities seems to work only for continuous change (the differentiable kind of growth), but all the interesting empirical applications have to do with catastrophes (in the mathematical sense), which is to say, nondifferentiable features of form change. On the whole there isn't even a language of empirical possibilities beyond the wholly abstract mathematical bestiary, let alone the empirical models that quantify their variability in practice. If something is invaginating, do the coordinates pertain to where it began or where it ends up? Pictures, biochemical and biophysical models, and statistics of all sorts will differ radically depending on the choice here. As one might imagine, everything about how subsequent computations proceeds is a function of this first decision about how to keep track of coordinates that are actually moving around or being created or destroyed over time. Nor are there good classification systems from which words such as "involute", "project", "infiltrate" arise as instances. These seem to be referring to dynamics, but in every case the apposite model is merely metaphorical. The collaboration of morphometrics and bioinformatics is needed to collect this vocabulary, sort it into bins by mathematical complexity, and start to work out the semantics of multiple overlapping descriptors that would be required.

3 A proposal

One tentative step toward the synthesis I'm advising would be an informal classification of the kinds of coordinate systems by which bioinformatic spatial representations will attach to a variety of structures. Recent work of mine, based in interesting navigations of Eve, extracted six generic prototypes. Each of them specifies a family of diffeomorphisms, but not of the general Michael Miller type, rather with some derivatives constrained along one or more curves or surfaces.

1. *Lineal* coordinate systems, such as apply to tendons, major nerves, or long bones, center on one lineal coordinate together with the corresponding series of normal planes, across which there may be no further orientability. These coordinate systems can be, but need not be, anchored by landmarks at one or both ends.

2. *Angular* coordinate systems, arising as constrained warps of coaxial pencils of planes. Statistics can proceed in terms of differences of the principal angle together with displacements in hemiplanar Cartesian systems (x, r) of the several sections.

3. *Cylindrical* coordinate systems, suited for such extended structures as the descending aorta, have one each of the three main coordinate types: lineal (position along the structure), radial (distance from a curving axis modeling the structure), and angular (with respect to some pertinent large-scale orientation—for the descending aorta, this is the anatomical mediolateral).

4. *Surfi cial* coordinate systems combine the standard approach of classical differential geometry—two tangential in-surface coordinates—with a surface normal distance. In practice, it helps if one of the tangential coordinates is highlighted.

5. *Spherical* coordinate systems are made up of "latitude, longitude, and altitude," as for the head of the femur or the globe of the eye. The meaning of "altitude" here may be absolute (as for the femoral head with respect to its socket) or may covary gently with latitude and longitude in some privileged coordinate system (e.g., for the myopic orbit, one with the pupil as a "north pole").

6. *Symmetrical* coordinate systems. The symmetry curve of a symmetric anatomical struc-

ture is a powerful new morphometric formalism: a curve "up the middle" of an extended structure together with a ruled surface of lines perpendicular to the ordinary tangent lines.

When the coordinate system is variable, landmarks not only *have* coordinates but also serve to anchor these coordinate systems, for instance, as terminations or branch points of axial structures, intersections of curves with surfaces or intersections of curves with curves on surfaces, centers of sufficiently small inclusions, extremes of 2D or 3D curvature (the INRIA school of extremal points), intersections of curves with the symmetry plane, etc. The new classification of landmark types that results (Bookstein et al., 2004), which emphasizes their typical origin in relations among curves and surfaces, has at last superseded that from my Orange Book of 1991.

4 Implications for bioinformatics

Our collective need is to articulate these and other approaches to anatomical variability with the great variety of other ontological terms and concepts by which the present and future bioinformatic literature will be formalized: specifically, to extend all other bioinformatic toolkits by a complementary semantics of modes by which published literature expresses anatomical information. In some research communities, for instance the people engaged in "human brain mapping," verbal descriptions of parts such as Brodmann areas have already been replaced by descriptions in terms of coordinates in a standardized system. My argument here is that this replacement does not go far enough for good developmental models, for literature standardization, for data mining across studies, or even for good morphometrics.

Typically, an article about molecules or mechanisms that happen to be based in organismal data will state something like the following: "sections were taken in the caudal half of the kidney, posterior to the calyx, aligned with the collecting tubules". Any bioinformatic representation of this information needs to recognize three different vocabularies here: the language of the FMA (parts of the body: kidney, calyx), the language of orientation ("aligned with", a term of art from morphometrics), and the language of histology ("tubules"). A search engine would need to know, for instance, that there is a conventional coordinate system for the kidney, and that tubules are part of the collection system; and it needs to know that tubules are lineal, so that being "aligned with" them requires two more anatomical specifications before the geometrical specification is completed. In some of these constructions, furthermore, more than one anatomical entity is invoked: for instance, "the aorta at the spring of the brachiocephalic trunk," or "the optic nerve midway along its length." Other semantic schemes apply to descriptions of dynamic processes, such as invagination or infiltration, and to static shape when described by evocative visual language: the tip of a bulge, the rim of a disk, the center of a convex blob.

All of these approaches lead to what the literature of morphometrics calls "deficient geometrical specifications," specifications in which some coordinates that would otherwise be crucial are missing: specifications of a full three Cartesian coordinates of location, for instance, or the full symmetric tensor of the spatial derivatives of locations, or the full 2×3 rotation matrix indicating the orientation of a section with respect to the organ's, or the body's, own canonical system. Whenever possible, the representation of the resulting literature retrieval should be in terms of these actual coordinates, together with their uncertainties. "The aorta at the spring of the branchiocephalic trunk," for instance, would be specified as a range of standardized coordinates in the standard template (Adam or Eve, perhaps)—this structure is a three-holed surface on either of those specimens—along with the corresponding exterior derivatives. Similar considerations apply for incompletely specified section planes.

5 Envoi

Such a system, combining anatomical and morphometric language in an ontology for multiscale positional specification in medicine, would greatly enhance the bridge linking bioinformatics to organismal biology. Any sequence-keyed or pathway-keyed bioinformatic fact is characterized by location in the normal (or abnormal) body in addition to any molecular or genetic properties it might encode. The extension required from the current state of morphometrics emphasizes local information about coordinate derivatives, information that was once put forward as a timely extension of morphometrics (Bookstein and Green, 1993) but was thereafter apparently forgotten except for Mardia et al. (2004). Now, with the current turn to bioinformatics, would be a good time to reopen the general issue of how informatic data bases attach to real, variable, sometimes even squirming organisms.

References

Bookstein, F.L. and Green, W. D. K. (1993). A feature space for edgels in images with landmarks. *Journal of Mathematical Imaging and Vision*, **3**, 231–261.

Bookstein, F.L.,Streissguth, A., Sampson, P., Connor, P., and Barr, H. (2002a). Corpus callosum shape and neuropsychological deficits in adult males with heavy fetal alcohol exposure. *NeuroImage*, **15**, 233-251.

Bookstein, F.L. Sampson, P.D., Connor, P.D., and Streissguth, A.P. (2002b). The midline corpus callosum is a neuroanatomical focus of fetal alcohol damage. *The Anatomical Record – The New Anatomist*, **269**, 162–174.

Bookstein, F.L.˙(2004) After landmarks. In press in in *Modern Morphometrics in Physical Anthropology* (D. E. Slice, ed.). Kluwer Academic Publishers, New York.

Bookstein, F.L. Schäfer, K., Mitteroecker. P., Gunz, P., and Seidler, H. (2004). The geometry of anthropometrics: a new typology of landmarks. *American Journal of Physical Anthropology*, **S38**, 66.

Flanders, H. (1963). *Differential Forms, with Applications to the Physical Sciences.* Academic Press, 1963.

Mardia, K.V., Kirkbride, J., and Bookstein, F.L.˙(2004). Statistics of shape, direction, and cylindrical variables. *Journal of Applied Statistics*, **31**, 465–479.