Int. J. Therm. Sci. (1999) 38, 653–663

How nature takes shape: extensions of constructal theory to ducts, rivers, turbulence, cracks, dendritic crystals and spatial economics

Adrian Bejan*

Department of Mechanical Engineering and Materials Science, 144A School of Engineering, Box 90300, Duke University, Durham, NC 27708-0300, USA

(Received 18 March 1999, accepted 20 March 1999)

Abstract — The constructal theory of the origin of geometrical form in natural flow (open) systems began with the discovery that, contrary to the established view, the tree network can be deduced from a single principle: the geometric minimization of resistance in volume-to-point flow. This article reviews a series of developments that extend the constructal law over naturally shaped flow phenomena other than the tree. Examples include the proportionality between width and depth in rivers of all sizes, the nearly round cross-sections of all blood vessels and bronchial passages, the dendritic shape of the snowflake, the pattern formed by cracks in a solid that shrinks upon cooling or drying (e.g., mud cracks), the onset and multiplication of rolls in Bénard convection, the transition (first eddy) and stepwise growth of all turbulent mixing regions, and the very existence of economics spatial structure (minimal cost routes between an area and one point). © Elsevier, Paris.

constructal therory / tree network / minimization of resistance / economic optimization / transition to turbulence / Bénard rolls / shape of the cross-sections

Résumé — Comment la nature prend forme : extension de la théorie constructale aux réseaux de canalisations, aux rivières, à la turbulence, à la convection naturelle, aux crevasses, aux réseaux dendritiques et à l'économie. La théorie constructale sur l'origine des formes géométriques dans les systèmes d'écoulement naturels commence avec la découverte du fait que, contrairement à un point de vue établi, les réseaux arborescents se déduisent d'un principe simple : celui de la minimisation géométrique de la résistance à l'écoulement à partir d'un point, dans un volume donné. Dans cet article, on passe en revue une série de développements qui permettent d'étendre la loi constructale, relative à la forme des écoulements naturels, à d'autres phénomènes, non arborescents. Les exemples présentés incluent la proportionnalité entre largeur et profondeur des rivières de toutes tailles, la forme approximativement circulaire des vaisseaux sanguins et bronchioles, la forme dendritique des flocons de neige, les motifs des craquelures qui apparaissent dans les solides en refroidissement ou lors du séchage, l'apparition, puis la multiplication des cellules de Bénard en convection, la transition vers la turbulence, puis sa croissance et, enfin, l'existence d'une structure spatiale des économies (itinéraire à coût minimum entre une région et un point). © Elsevier, Paris.

théorie constructale / réseaux arborescents / minimisation des résistances / optimisation économique / transition vers la turbulence / convection naturelle / forme des sections droites

1. CONSTRUCTAL THEORY OF ORGANIZATION IN NATURE

This article is an invitation to think freely about a phenomenon that is so prevalent that it is being taken for granted: the macroscopic shapes and structures that generate themselves everywhere in nature [1]. It is an invitation to think about the great puzzle that has been with us from the beginnings of science: "From what principle can geometrical form be *deduced*?" Democritus (c. 460–c. 370 BC) attributed natural geometrical form to "chance and necessity." The doctrine of chance (nondeterminism) has stayed with us ever since, not as an explanation of natural form generation but as an admission of our own inability to predict it.

* J.A. Jones Professor of Mechanical Engineering

abejan@duke.edu

This paper is a follow-up to a communication presented at the ECOS Conference, which took place in Nancy (France) from 8 to 10 July 1998.

Let us start with a few empirical observations. Geometric form is generated in natural systems that are internally 'alive' with flows and driving gradients (e.g. temperature and pressure). Such systems are not in equilibrium internally. Second, the geometric forms that our minds recognize and sort out are not many. Three shapes cover most of the world that is around us and inside ourselves: the tree-shaped flow network, the round shape of the cross-section of duct flow, and the watermelon-slice shape of the cross-section of open channel flow. Third, natural systems that have the same shape are not identical. For example, two bronchial trees are never identical. Similarly, two cuts made across a blood vessel never reveal the perfect, mathematical circle. The point is that when presented with one image from the endless diversity of natural flow shapes, the mind knows this image and categorizes it as a tree, round or watermelon-slice shape. The tree may be hard to describe, but when we see it we know it, and we call it 'tree'.

These few natural shapes are truly everywhere, in both animate and inanimate flow systems. If a single principle — a simple statement — is responsible for the generation of billions and billions of such shapes, then that law manifests itself everywhere, and bridges the gap between the scientists' physical and biological fields of vision.

Constructal theory is about the physics principle from which geometric form in natural flow systems can be deduced. This line of inquiry began accidentally in engineering, with a 1997 analytical paper on the conductive cooling of a small electronic package (a heat generating volume) by using a point-size heat sink [2]. I proposed the fundamental problem of how to connect one point to an infinity of points. In the electronic package, the volume is fixed, and the heat generating material has a low thermal conductivity (k_0). A small amount of a second material — one with much higher thermal conductivity (k_p) — is to be distributed through the k_0 material such that the overall volume-to-point resistance is minimal. The accidental discovery is that by invoking consistently a single principle — the minimization of volume-to-point resistance — we find that the k_p material fills the links of a tree network that is completely deterministic. In this solution even the integer 2 (bifurcation, pairing) is an optimization result, not an assumption.

These features can be rediscovered in the tree networks that occur naturally (e.g., fluid flow, electricity, streets) [1]. Constructal theory and its first applications were reviewed by Bejan and Tondeur [3] in the alternate context offered by the principle of equipartition [4] or optimal allocation of hardware, as a route to thermodynamic optimization subject to constraints. A progress report on constructal tree networks is given in section 8. The following is a brief review of several additional directions in which constructal theory has been extended.

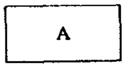
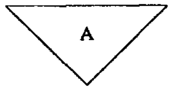
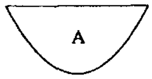
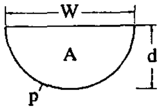
2. NATURAL CHANNEL AND DUCT CROSS-SECTIONS

To predict the shape of the river cross-section may appear difficult, if we think of the real world and the many uncertainties of river flow and channel development. Turbulent flow, secondary flow (cross-circulation), meanders, bed erosion, sediment transport, time, and the geological characteristics of the terrain all play important roles, which have been amply documented in the river morphology literature. These uncertainties are responsible for significant variations in channel cross section from one river to another, or along the same river. They can also be invoked when observing that the ratio between river width (W) and the maximum depth (d) is not a constant: there is some scatter in the W/d data.

The classical explanation for the shape of the channel cross section recognizes the complicated, developing nature of the river bed [5]. An equilibrium was envisaged between the forces acting on a ground particle at rest on the bottom of the channel. This equilibrium condition, however, is insufficient and must be complemented by ad-hoc assumptions concerning the distribution of shear stress (or bottom velocity) with depth. After numerical integration, this approach yields a bottom shape that is roughly sinusoidal, in which W and d are two undetermined constants. In sum, the equilibrium theory of the river bed does not explain the proportionality between W and d.

This problem becomes much easier if we focus on an open channel cross section with an upper straight segment, W (the shear-free surface), and the rest of the perimeter, p (the bottom), as shown in the *table.* The shape of the bottom curve of length p is not specified. The channel cross-sectional area A, i.e., the

TABLE
Optimized cross-sectional shapes of open channels.

	Optimal Shape	$(W/d)_{opt}$	$p_{min}/A^{1/2}$
Rectangle		2	2.828
Triangle		2	2.828
Parabola		2.056	2.561
Circle		2	2.507

area enclosed by the W segment and the p curve, is fixed. When the flow is turbulent and the bottom sufficiently rough, the skin friction coefficient along the bottom is nearly independent of flow rate. This means that the minimization of the flow resistance is equivalent to the statement: find the cross-sectional shape that has the minimum perimeter p, which along with W encloses the area A.

The solution — the optimal image — has two parts, i.e., two degrees of freedom. First, the optimal shape of the p curve is delivered by variational calculus, and it is the arc of a circle [1]. There is an infinity of such cross sections, depending on the size of the ratio $W{\cdot}A^{-1/2}$. In one extreme, when this ratio is zero, the cross section is a complete disc, and represents the well known solution for the duct (blood vessel, bronchial tube). The second degree of freedom is represented by $W{\cdot}A^{-1/2}$, or by the ratio W/d. The optimal value of the latter is 2. In conclusion, the optimal shape of the channel cross-section can be derived from the same principle of global optimization subject to constraints as in the other natural flow shapes reviewed in this article.

It is interesting that in the optimal shape (half disc) the river banks sink vertically downward into the water, and are likely to crumble under the influence of gravity and erosion (drag on particles). This effect will decrease the slope of the river bed near the surface and, depending on the bed material, it will increase somewhat the slenderness ratio W/d. There remains plenty of room for the classical equilibrium theory of the river bed, in fact, its territory remains intact. Equilibrium theory begins where constructal theory leaves off.

If we do not know the variational calculus solution for the optimal bottom shape (arc of circle), we can still assume a shape (e.g., rectangle, triangle) and optimize its W/d ratio for minimum resistance. The rightmost column of the *table* shows that in practical terms these alternate shapes have nearly the same resistance as the best shape. This high level of agreement accounts for the scatter in the W/d data on river bottom profiles, that is, if global thermodynamic performance is what counts, not local details. Yes, there is uncertainty in the actual shapes that we see in nature. Important is that there is very little uncertainty in anticipating global characteristics such as shape (e.g. round vs. tree), optimal performance, and basic mechanism. Additional support for this view is provided by the billions and billions of internal ducts found in plants and animals. Ducts with imperfections (flat spots) perform almost the same as purely round ducts [1].

3. TURBULENCE: THE FIRST, SMALLEST EDDY

In every natural tree example [1] the flow has the property to develop structure when it can exist in two regimes, not one. Each flow path starts from the elemental volume with a portion with high resistivity (diffusion), and continues with several portions with low resistivity (streams) at larger scales. Turbulent flow is notorious for combining the same two regimes — viscous diffusion and streams (eddies) — therefore, it must be covered by the constructal law of structure generation in nature. To see how, let us consider the sudden shearing motion between two semi-infinite fluid reservoirs, with the velocity U_∞ measured between them. We see this motion better in a frame that rides at half speed along the interface ($y = 0$): The upper fluid rides to the right at $U_\infty/2$ and the lower fluid moves to the left with the same speed, as shown in *figure 1.*

Immediately after $t = 0$, the interface is thickened by a laminar shear layer with error-function profiles on both sides of the $y = 0$ plane. The instantaneous thickness of this layer (D) is a classical result of Stoke's first problem. The effective distance of viscous penetration (thickening, mixing) is marked by the knees in the velocity profile, which are located at $y \approx \pm 2\,(\nu\,t)^{1/2}$, where ν is the kinematic viscosity of the fluid. In the present case, viscous diffusion propagates on both sides of the $y = 0$ interface, therefore $D \approx 2\,y \approx 4\,(\nu\,t)^{1/2}$.

The question suggested by the constructal principle is this: how can the flow (velocity nonuniformity) spread itself over the entire space in the shortest time possible? The laminar regime is effective only in the beginning when $\mathrm{d}D/\mathrm{d}t$ is large. As time increases, the diffusive swelling of the mixing region slows down because $\mathrm{d}D/\mathrm{d}t$ decreases as $t^{-1/2}$.

To maintain the highest rate of growth possible, the system searches for a second regime. In shear flow, the second regime is eddy formation, or shear layer roll up. The vertical motion of this organized motion is constant ($U_\infty/2$), because the peripheral speed of the roll is set by the motion of the two reservoirs. During one roll-up the shear region swells to a thickness equal to the diameter of the first roll. The latter is approximately equal to $\lambda_\mathrm{B} \sim 2\,D$, where λ_B is the wavelength of buckling (or neutral) deformation of the viscous shear layer of thickness D [6, 7]. In other words, the mixing thickness increases from D to $2\,D$ during the time required by one roll-up, $t_\mathrm{B} \sim \lambda_\mathrm{B}/(U_\infty/2) \sim 4D/U_\infty$. Comparing these competing regimes we conclude that the rate of lateral growth through eddy formation is $(2\,D - D)/t_\mathrm{B} \sim U_\infty/4$, whereas the viscous swelling rate is $\mathrm{d}D/\mathrm{d}t \sim 2\,(\nu/t)^{1/2} \sim 8\,\nu/D$. The most rapid growth occurs when diffusion is followed by eddy formation at a time when the viscous growth rate is just outpaced by the eddy growth rate — that is, when $U_\infty/4 \sim 8\,\nu/D$, or more appropriately:

$$D\,U_\infty/\nu \sim O(10^2) \qquad (1)$$

This *local Reynolds number* criterion predicts the transition to turbulence in all known configurations [7], where the classical data are a collection of many transition constants that vary in order of magnitude from a Reynolds number of order 10^2 in jets and wakes,

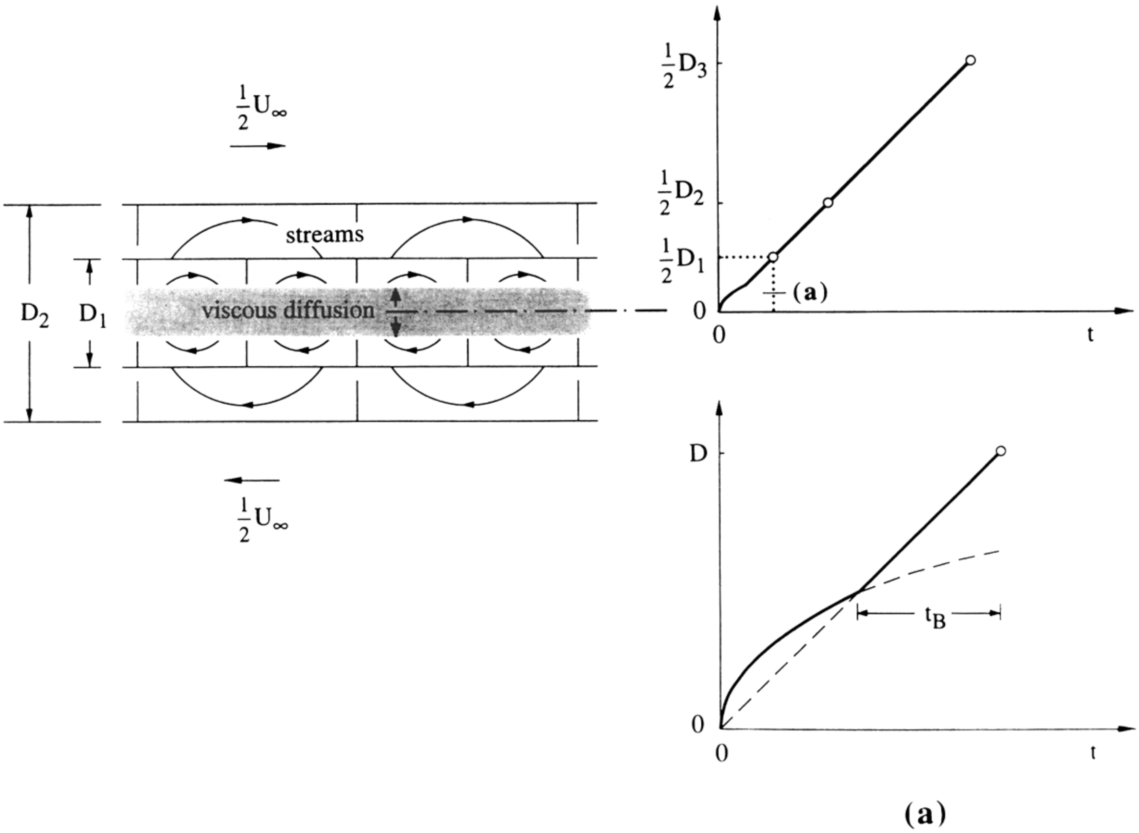


Figure 1. The maximization of spreading rate (a), the first eddy, and the subsequent stepwise growth of a turbulent shear-flow mixing region [1].

to a Rayleigh number of order 10^8 in natural convection vertical boundary layer flow. The latter is anticipated by the theoretical statement that at transition the Reynolds number must be of order 10^2, where the Reynolds number is based on the local thickness and local longitudinal velocity of the flow.

The first roll-up is the smallest eddy — the elemental volume of the subsequent constructs that will make up the turbulent flow field. The elemental volume is ruled by a time balance: the time of viscous diffusion matches the time of rolling once. This balance is analogous to the equipartition of driving force [3, 4] in the elemental volume of constructal trees [1]. The size of the smallest eddy is such that its Reynolds number based on diameter and peripheral velocity is of order 10^2. This purely theoretical result establishes order in contemporary fluid mechanics, where the smallest eddy is thought to have a Reynolds number of order 1.

Dramatic support for the theoretical size of the first eddy is offered not only by the overwhelming record on turbulent flows, but also by the massive record of observations on the swimming of fish. Small fish flap their fins and hold their bodies straight when the Reynolds number is less than $O(10^2)$. When the Reynolds number exceeds $O(10^2)$, fish swim by undulating their bodies. Now we know why. The swimming of fish visualizes the shape and structure of the surrounding medium. The fish had to acquire the shape of its medium in order to minimize its own resistance to flow. This is the first time that the transition in fish swimming has been predicted based on pure theory, and is a most unexpected and rewarding result of constructal theory.

Beyond the first eddy formation event (D_1), the flow continues to expand in steps — in assemblies, or constructs — by rolling and forming eddies. Each step leads to the doubling of the mixing region (cf. $\lambda_B \sim 2D$). Viscous diffusion does not have time to act (to compete with rolling) over larger distances such as D_2, D_3, and so on. The outer boundaries of the mixing region are extended now by the second flow regime: streams (eddies). In a frame of reference attached to one of the fluid reservoirs, the mixing region grows stepwise as a stack of geometrically similar building blocks [6, 7]. In a volume averaged description the mixing region appears to have the shape of a wedge, i.e., a constant growth rate. The analogy with the other structures deduced based on constructal theory is now complete.

4. BÉNARD CONVECTION

Consider a horizontal layer of single-phase fluid heated from below, which is characterized by the thickness H and the bottom excess temperature is $\Delta T = T_h - T_c$. In line with the access-optimization principle of constructal theory, we search for the fastest (most direct) route for heat transfer across the fluid layer [1, 8]. To start with, the classical solution for time-dependent thermal diffusion near a wall with a sudden jump in temperature (ΔT) is $(T - T_c)/\Delta T = \text{erfc}[y/2\,(\alpha\, t)^{1/2}]$, where T_c is the far-field temperature in the fluid. The effect of the temperature jump is felt to the distance $y/2\,(\alpha t)^{1/2} \sim 1$, which represents the knee in the temperature profile. The time needed by this heating effect to travel by thermal diffusion the distance H is $t_0 \sim H^2/(4\,\alpha)$. The time t_0 corresponds to the heating of the entire layer ($y \sim H$). The factor 4 in the denominator of the t_0 expression arises from the geometry (shape) of the time-dependent temperature profile.

Pure conduction continues to be the preferred heat-transfer mechanism, and the fluid layer remains macroscopically motionless as long as H is small enough that t_0 is the shortest time of transporting heat across the layer. The alternative to conduction is convection, or the channeling of energy transport on the back of fluid streams. The question is whether the convection time (t_1) across the layer H is shorter than t_0. The convection time is $t_1 \sim H/v$, where v is the vertical velocity of the fluid (the peripheral velocity of the roll).

To evaluate the v and t_1 scales, we rely on scale analysis [7]. First, we note that the effective diameter of each roll is of order H, but smaller, for example, $H/2$. When the roll turns, an excess temperature of order $\Delta T/2$ is created between the moving stream and the average temperature of the fluid layer. This excess temperature induces buoyancy (modified gravitational acceleration) of order $g\,\beta\,\Delta T/2$. The total buoyancy force that drives the roll is of order $(g\,\beta\,\Delta T/2)\,\rho\,(H/2)^2$. When the Prandtl number is of order 1 or greater, the driving force is balanced by the viscous shearing force $\tau\,H/2$, where the shear stress scale is $\tau \sim \mu v/(H/4)$. The force balance buoyancy–friction yields the velocity scale $v \sim g\,\beta\,\Delta T\,H^2/(16\,\nu)$ and the corresponding convection time scale $t_1 \sim 16\,\nu/(g\,\beta\,\Delta T\,H)$.

To see the emergence of an opportunity to optimize the geometric features of the flow pattern, imagine that H increases. As the system grows, the thermal diffusion time t_0 increases in accelerated fashion, whereas the convection time t_1 (a property of the H system, even if quiescent) decreases monotonically. Setting $t_1 \lesssim t_0$ we find that the first streams occur when $Ra_H \sim O(10^2)$, where $Ra_H = g\,\beta\,\Delta T\,H^3/(\alpha\,\nu)$ is the Rayleigh number. The exact solution for this critical condition is $Ra_H = 1708$; in other words, $Ra_H = O(10^3)$. The factor-of-6 error in the result of scale analysis is understandable (and unimportant) because it can be attributed to the imprecise geometric ratios (factors of order one) introduced en route to determining the t_0 and t_1 scales. What is important is that the predicted critical Ra_H is a constant considerably greater than 1. This constant is a conglomerate of all the geometric ratios of the roll-between-plates configuration. Had we neglected the geometric reality of how the rolls fit, or the geometric fact that 4 belongs in the denominator of the t_0 expression, we would have obtained only $Ra_H \sim 1$, i.e. the correct dimensionless group but not the fact that the critical Ra_H number represents *geometry* (structure).

When convection occurs, there are two heat-transfer mechanisms, not one. Each roll characterized by $t_0 \sim t_1$ is an elemental system in the sense of constructal theory. The equipartition of time $t_0 \sim t_1$ is the analog of the equipartition of temperature drop across an optimized element of the heat generating volume of constructal theory. Conduction, or thermal diffusion, is present and does its job at every point inside the elemental volume. Superimposed on this volumetric heat flow is an optimal pattern of convection 'streets' that channel the imposed heat current *faster* across H.

The usual terminology for 'faster' in the field of heat transfer is to say that the onset of convection is followed by an increase in the overall Nusselt number, $\overline{Nu}_H = \overline{q''}\,H/(k\,\Delta T)$. If we fix the uniform heat flux q'' in Bénard convection, we see again that the optimization of the heat flow pattern at the elemental level leads to a smaller overall ΔT, and thus a larger $\overline{Nu}_H$. The geometric minimization of the temperature difference across H continues to manifest itself as H (or Ra_H) increases, as convection becomes more intense. In this case, geometric optimization means the selection of the number of rolls that fill a layer of fixed horizontal dimension L, or the selection of the roll aspect ratio. This principle of natural optimization of the flow geometry is known as the *Malkus hypothesis* and was proposed heuristically in the usual context of maximizing $\overline{q''}$ when ΔT is imposed [9].

Constructal theory has also been extended to Bénard convection in porous layers saturated with fluid in the Darcy flow regime [8]. The geometric structure and heat transfer rate in Bénard convection at Ra_H values higher than critical has been predicted in amazingly simple and direct terms (analytically) by continuing to apply the constructal principle of access optimization for internal currents [8]. The analytical method is illustrated for crack patterns in § 6.

5. DENDRITIC CRYSTALS

The dendritic crystals that form during rapid solidification are another wide class of naturally ordered solid shapes. In 1611 Kepler drew attention to the shapes, numbers, and geometric similarities exhibited by snowflakes [10]. In this century, the study of dendritic crystals has grown into a major field that deals mainly with two aspects: the shape and the growth of dendritic crystals. Not questioned was the *necessity* of the dendrite. Why should the needles be necessary?

Consider the solidification of a single-component substance, and assume that the liquid and solid phases have the same density. This means that there is no liquid motion in the vicinity of the solidification front: what flows is heat, which starts from the solidification front and diffuses into the liquid. The solid phase is isothermal at T_s. Sufficiently far from the solidification front the liquid is metastable (subcooled) at the temperature $T_\infty (> T_s)$. Solidification starts at $t = 0$.

Why Plane? Kepler asked the first fundamental question about the geometry of the snowflake: why six needles? This question was answered after the geometry of the molecular arrangement of ice became known. An equally important question is this: why is the snowflake plane? This question cuts to the heart of the meaning of 'rapid' solidification.

The classical solutions for unidirectional solidification in plane, cylindrical, and spherical geometries show that in all cases the solid thickness (R) and the thickness of thermal diffusion into the liquid (r_l) increase as $(\alpha t)^{1/2}$, where α is the thermal diffusivity of the liquid. The three geometries are quite different with regard to their ability to fill the space with solid: specifically, (solid volume)/(total volume) $\sim (R/r_l) < 1$ (plane), $\sim (R/r_l)^2 << 1$ (cylinder) and $\sim (R/r_l)^3 <<< 1$ (sphere). Clearly, the most effective arrangement for solidifying a volume in the shortest time is the planar one. That snowflakes are plane is overwhelming evidence that the maximization of the speed of volumetric solidification is an integral part of the nature of the process.

Why needles? The preceding observation brings us to the most important problem, which is that of predicting the necessity of the dendrite. Let us follow, in time, the growth of the snowflake beginning with its birth ($t = 0$) at a point-size nucleation site (*figure 2*). Molecular structure and surface-tension-based stability arguments indicate that ice needles will begin to grow with the velocity U (constant) in the six directions shown. But this is only one heat transfer regime of the solidification process that just started. At the same time ($t = 0$), the temperature of the nucleation site jumped to T_s ($> T_\infty$) and triggered a spherical wave of thermal diffusion (warmth) the radius of which increases as $r_l \sim 2(\alpha t)^{1/2}$. The initial speed of propagation of the liquid heating effect is infinite—that is, larger than any constant speed (U) that the tip of the needle might have. In time, the speed dr_l/dt decreases as $t^{-1/2}$ and is eventually overtaken by U.

Figure 3 shows how the warmed liquid sphere and the needle length grow in time. A critical time t_c is reached when the needle length L overtakes the radial length scale of the warmed liquid, $U t_c \sim 2(\alpha t_c)^{1/2}$, which

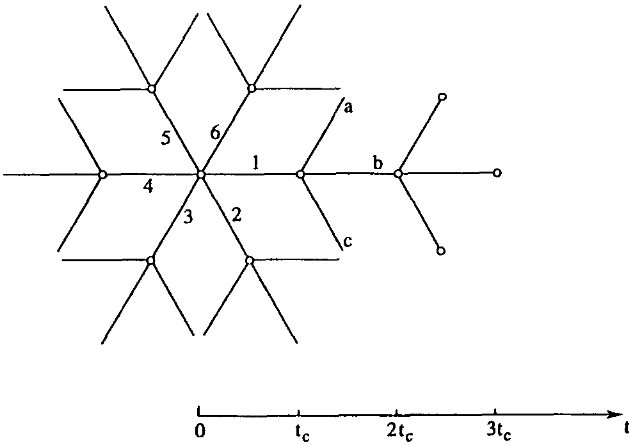
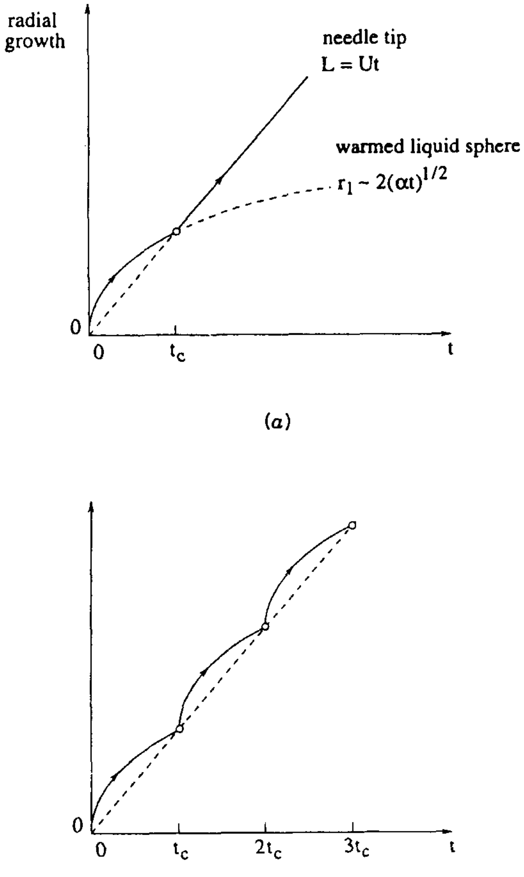


Figure 2. The formation of new needles after each time interval t_c, as the repeated manifestation of the mechanism shown in *figure 3* [1].



(a)

(b)

Figure 3. The simultaneous growth of the needle and the warm liquid sphere, and the time interval t_c after which the process is repeated.

yields $t_c \sim 4\alpha/U^2$. Note that *figure 3a* is completely analogous to *figure 1*. At times slightly greater than t_c, the needle of length $U t_c$ just sticks its tip out of the warmed liquid sphere. The tip is once again surrounded by isothermal subcooled liquid from the half-space that lies in front of it. The situation at $t \sim t_c$ is the same as that at $t = 0$, except that the new nucleation site (the needle tip) can send new needles only forward, because the trailing half space is already warmed and/or solidified. In conclusion, at $t \sim t_c$, each tip serves as nucleation site for three forward-leaning fresh needles. There is no difference between the ages of these fresh needles. Because of the 60° angular symmetry, however, the middle needle (b) in *figure 2* looks like a continuation of the original needle, and, consequently, the other two needles (a, c) look like 'branches'.

From $t \sim t_c$ until $t \sim 2t_c$, the new generation of needles experiences the 'growth inside the warm liquid sphere' process, that we saw between $t = 0$ and $t = t_c$. One difference is that the liquid spheres of the side needles (a, c) interfere eventually with the spheres of the side needles of the adjacent original directions. Consequently, at $t \sim 2t_c$ all the side needles become suffocated by the warm (saturated) liquid environment, and their needle-like growth ceases. Each middle needle (b) however, pierces its warmed liquid sphere and generates another group of three forward-leaning fresh needles. This third generation and its fully grown version at $t \sim 3t_c$ are illustrated in *figure 2*, which is based on repeating the argument of *figure 3a* three times, as shown in *figure 3b*.

With this, the task of predicting the architecture (existence) of dendrites has been accomplished. Needles are necessary for the same reason that eddies are necessary: to provide internal paths such that entire volumes approach internal equilibrium in the shortest time possible.

Imagine that you are a visitor from another planet, someone who knows biology but not thermophysics. You may describe *figure 2* as follows. The organism is born at the time $t = 0$, and its innermost morphology (e.g. the number 6) is a reflection of information stored at the molecular level. The organism grows — that is, expands into its surroundings. The growth is very fast when the organism is young, and it slows down with age. There comes a time — the time of death, t_c — when the organism becomes disposable. The fossil (solid dendrite) is suggestive of the flow that was present in the living organism (heat flow). Continuity is assured in the form of several offspring, which repeat the life cycle of the original organism. The offspring may or may not be attached to their ancestor. In sum, the periodic regeneration and multiplication of the organism is an expression of the natural tendency toward internal geometry for optimal (or fastest) access for internal currents.

6. CRACKS IN SOLID SURFACES

The formation of cracks in solids is an old and busy field that so far has been dug mainly by materials scientists, physicists, and chemists. The challenge that persists is to predict the origin of such patterns — that is, to explain why they are necessary. During the past decade it has become fashionable to describe cracking patterns in terms of fractal images. This tool is pleasing, but not predictive.

Let us think freely about the most common example of patterned cracks. Wet soil exposed to the sun and the wind becomes drier, shrinks superficially, and develops a network of cracks. The loop in the network has a characteristic length scale. The loop is round, more like a hexagon or a square, not slender. The loop is smaller (i.e., cracks are denser) when the wind blows harder — that is, when the drying rate is higher.

These unexplained characteristics of mud cracks are hints that their pattern is another natural occurrence of access optimization: the maximization of the mass transfer rate from the system (wet soil) to the ambient, or the minimization of the overall drying time. In view of the analogy between mass transfer and heat transfer, we can explore this theoretical route by considering the thermal analog sketched in *figure 4*. A one-dimensional solid slab of thickness L is initially at the high temperature T_H, and has the property to shrink upon cooling. The coolant is a single-phase fluid of temperature T_L. The question is how to maximize the thermal contact between the solid and the fluid or how to minimize the overall cooling time. The obvious 'design' is to allow the fluid to flow through the solid. In *figure 4* the cracks are spaced uniformly, but their spacing (R) is arbitrary. The channel width (D) increases in time, as each solid piece (R) shrinks. The fluid is driven by

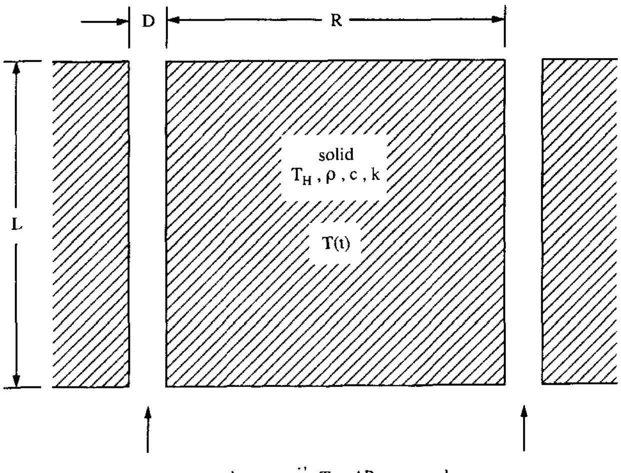


Figure 4. Channels in a shrinking solid cooled by single-phase fluid [1, 11].

the pressure difference ΔP, which is maintained across the solid thickness L. The imposed ΔP is an essential aspect of the channel spacing selection mechanism. For example, in the air cooling of a hot solid layer the scale of ΔP is set at $(1/2)\,\rho_f U_\infty^2$, where ρ_f and U_∞ are the density and free-stream velocity of the external air flow.

To examine the effect of the channel spacing R on the time needed for cooling the solid, we consider the two asymptotes $R \to 0$ and $R \to \infty$. The approach is the same as in the geometric optimization of electronic packages [7]. In other words, electronic packages emerge as patterns of heat-generating blocks separated by optimal cooling channels for the same reason that optimal patterns of cracks occur in nature. This observation reinforces the commonality of natural and man-made patterns under constructal theory.

When the number of channels per unit length is large, the spacing R is small and so is the eventual shrinkage that is experienced by each R element. This means that when $R \to 0$ we can expect $D \to 0$ and laminar flow through each D-thin channel, such that the channel mass flowrate is $\dot{m}' = \rho_f\, D\, U \sim \rho_f\, D^3\, \Delta P/(\mu\, L)$. In the same limit, R is small enough so that the solid conduction is described by the lumped thermal capacitance model. The solid piece R is characterized by a single temperature T, which decreases in time from the initial level T_H to the inlet temperature of the fluid T_L. This cooling effect is governed by the energy balance $\rho\, c\, R\, L\, (\mathrm{d}T/\mathrm{d}t) = -q'$, where ρ and c are the density and specific heat of the solid. The cooling effect (q') provided by the flow through the channel is represented well by $q' = \dot{m}'\, c_p\, (T - T_L)$, where c_P is the specific heat of the coolant. We obtain the order of magnitude statement $\rho\, c\, R\, L\, (\Delta T/t) \sim \dot{m}' c_p\, \Delta T$, where ΔT is the scale of the instantaneous solid excess temperature $T - T_L$. Finally, by using the $\dot{m}'$ scale, we find the cooling time scale

$$t \sim \frac{\rho\, c}{\rho_f\, c_p}\, \frac{\mu\, R\, L^2}{D^3\, \Delta P} \qquad (R \to 0) \tag{2}$$

In the opposite limit, R is large and the shrinkage (the channel width D) is potentially very large — in proportion to R. The fluid present at one time in the channel is mainly isothermal at the inlet temperature T_L. The cooling of each solid side of the crack is ruled by one-dimensional thermal diffusion into a semi-infinite medium. The cooling time in this regime is the same as the time of thermal diffusion over the distance R,

$$t \sim \frac{R^2}{\alpha} \qquad (R \to \infty) \tag{3}$$

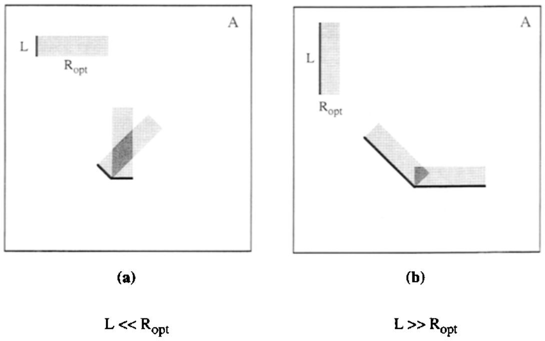
where $\alpha = k/(\rho c)$, and k is the thermal conductivity of the solid. To summarize, in the limit $R \to 0$ the cooling time is proportional to R/D^3 or R^{-2}, because we expect the proportionality $D/R \sim \beta\, \Delta T << 1$, where $\Delta T \sim T_H - T_L$, and β is the coefficient of thermal contraction of the solid. In the opposite limit, $R \to \infty$, the cooling time is proportional to R^2. Put together, these proportionalities suggest that the cooling time possesses a sharp minimum with respect to R or the channel density. Intersecting the two asymptotes we find that the optimal crack distance (R_{opt}) for fastest cooling is of the order of:

$$R_{\mathrm{opt}} \sim \left[\frac{k}{k_f}\, \frac{\alpha_f\, \nu\, L^2}{U_\infty^2\, (\beta\, \Delta T)^3} \right]^{1/4} \tag{4}$$

This result is promising for two fundamental reasons, in addition to the practical aspect of knowing how to extract heat or mass from the heart of a solid in the fastest way possible. One reason is that the optimal crack distance decreases as the external pressure (or flow) is intensified. This effect is in accord with numerous observations that mud cracks become denser when the wind speed increases. This result, in association with the theoretical view that natural cracks occur such that the cooling speed is maximized, is the first time that the effect of wind speed on crack density is predicted. The second reason is that the R_{opt} result predicts a higher density of cracks (a smaller R_{opt}) as the solid excess temperature ΔT increases. This trend too is in agreement with observations, and it is being predicted for the first time.

An important geometric aspect of the R_{opt} scale is that the optimal distance between consecutive cracks must increase as $L^{1/2}$. This is relevant to predicting the length scale of the lattice of vertical cracks formed in a horizontal two-dimensional surface cooled (or dried) from above, under the influence of external forced convection. As the air flow direction changes locally from time to time, and since the material (its graininess) is such that cracks may propagate in more than one direction, we arrive at the problem of cooling a two-dimensional terrain (area A, when seen from above) with cracks of length L and associated area elements of width R_{opt}.

Figure 5 shows the two extremes in which L may find itself in relation to R_{opt}. First, when L is considerably shorter than R_{opt} it is impossible to cover the area A



(a)

$L << R_{opt}$

(b)

$L >> R_{opt}$

Figure 5. The two extremes in covering a two-dimensional solid (A) with cracks (L) and optimally cooled volume elements ($L \times R_{\mathrm{opt}}$) [1, 11].

exclusively with patches of size $L \times R_{\text{opt}}$. The reason is that when two cracks of length L are joined at an angle, the elemental area ($\sim L^2$) trapped between them is too small to accommodate the amount of ideally cooled solid material. When L is considerably longer than R_{opt}, any lattice of cracks will fail to cover the area A completely. Now the trapped elemental area ($\sim L^2$) is considerably larger than the amount of ideally cooled solid ($\sim L\,R_{\text{opt}}$). This means that most of the interior of the area element of size L^2 would require a cooling time that is considerably longer than the minimum time determined in the preceding analysis.

In conclusion, to cool the entire solid (A) in the fastest way possible is to cover the A cross section with $L \times R_{\text{opt}}$ elements, in which $L \sim R_{\text{opt}}$. The optimal pattern is one with relatively 'round' or 'square' loops, not slender loops. Combining $L \sim R_{\text{opt}}$ with the R_{opt} expression we find the optimal length scale of the loop in the network of cracks that will minimize the cooldown time: $R_{\text{opt}} \sim (\alpha_f\, \nu\, k/k_f)^{1/2}/[U_\infty(\beta\,\Delta T)^{3/2}]$. Once again, in agreement with observations, we see that the lattice length scale R_{opt} must decrease as the wind speed and the initial excess temperature increase.

7. CLUMPS OF SOLID SPREADING THROUGH A FLUID MEDIUM

As an invitation in yet another direction of inquiry, let us ask why solid matter travels in large clumps through a fluid medium [1]. Why does it not travel as a swarm of very small granules? Why do celestial bodies form in time? Consider two granules — for example, two spheres of diameter D_1 and density ρ, such that their total mass is $m = 2\,\rho\,(\pi/6)\,D_1^3$. The drag force felt by each ball is $F_D = C_D\,(\pi/4)\,D_1^2\,(1/2)\,\rho_f\,U^2$, where r_f is the density of the medium (e.g., gas) that fills the space, and U is the relative velocity between ball and medium. For simplicity, assume that the Reynolds number is sufficiently large such that the drag coefficient C_D is almost constant and of order 1. In conclusion, the drag force experienced by the total mass m is $F_1 = 2\,F_D = (\pi/4)\,C_D\,D_1^2\,\rho_f\,U^2$.

Can the two masses reduce their resistance to travel? In other words, can the solid spread faster and farther through the fluid medium? Yes. Two balls fused into one larger ball encounter a smaller resistance than when they travel separately. Mass conservation dictates that the diameter of the larger ball is $D_2 = 2^{1/3}\,D_1$. The drag force on this larger ball, $F_2 = C_D\,(\pi/4)\,D_2^2\,(1/2)\,\rho_f\,U^2$ is sensibly smaller than in the original configuration, $F_2/F_1 = 2^{-1/3} = 0.794$. Given enough time, two neighboring masses should coalesce into a larger mass.

8. THREE-DIMENSIONAL TREE NETWORKS AND ANGLED TRIBUTARIES

The key aspect of the work reviewed in § 1–7 is the enormous number of natural examples, and the surprising diversity of the geometric flow structures that can be predicted based on constructal theory. I am sure that even more surprises are in store, that is, if we are willing to take an unbiased look at some of the unexplained and unquestioned forms of natural organization. The fields of physiology and geophysics are covered by such forms.

The reviewed work 'extended' constructal theory beyond the volume-to-point flows (tree networks) that got the idea started. The trees were deduced in two dimensions, as paths of minimum resistance to volume-constrained flows. The work on tree networks continued [12–15] with the objective of generalizing and communicating it to the physics community.

Several new developments on constructal trees are worth noting. The optimization of access was demonstrated in three dimensions, by minimizing the resistance to fluid flow [12], and by minimizing the time of travel between an entire volume and one point [13]. The latter also showed that it is possible to optimize the angles between tributaries and collecting path (street) in each new assembly. In the original work [1] on two-dimensional heat trees, fluid trees and street trees, the tributaries were assumed perpendicular to each collecting stream. It was shown through computer-based optimization that the angle optimization has only a minor effect on the minimized overall resistance of the system [15].

The initial work [1] was also based on the assumption that the flow can exist in two regimes, one with high resistivity at the elemental level, and the other with a considerably lower resistivity in the collecting paths of the constructs. The generalization that was just communicated [13, 14] is based on allowing any number of flow regimes, provided that number is greater than 1. For example, in the minimization of travel time between a volume and one point [13] the unspecified flow regimes are the unspecified speeds of travel along the central streets of the constructs. In the minimization of flow resistance between a heterogeneous porous medium and one point [14] the unspecified flow regimes are represented by the unspecified Darcy-flow permeabilities of the central paths (e.g., cracks) of the constructs. If the given volume can be covered with an assembly of order n, then the number of different flow regime (resistivities) is $n + 1$.

These theoretical developments and the older work on volume-to-point flows [1] remind us that a natural tree network is more than a 'stick drawing' that connects a root point with an area or volume. First, most of the tree is 'empty', without visible links (branches, tributaries). You can see through the tree. Second, links cannot be smaller than a characteristic length scale:

the number of stages of coalescence is *finite*. Third, the links that are closer to the root point are thicker. Fourth, the thicker links bifurcate or, when seen coming from the other direction, they coalescence into pairs. The integer 2, which stands for bifurcation or pairing (or dichotomy), is a defining characteristic of natural tree networks at higher levels of coalescence.

It is worth repeating that the constructal theory of volume-to-point flow [1, 2] began with the optimization of the smallest volume of known size, where volumetric flow ("diffusion") coexists with the first organized flow: the first channel, duct, or rivulet. This observation is particularly important in river morphology. It is hard to see such small "first rivulets" in nature, or even in controlled rain-erosion experiments. This is one reason why it has become fashionable to simulate drainage networks based on fractal algorithms, i.e., to assume that certain (postulated) similarity rules repeat themselves *ad infinitum* all the way to river links of size zero. In reality, the finite-size elemental areas and the birth of the first rivulets can be watched "live" in drinking cups coated with wet sediment from unfiltered coffee.

The common features of natural tree networks refute the notion that such patterns are fractals. Recall that the infinite sequence of fracturing steps is the defining statement of a fractal (ref. [17], p. 15): the noneuclidean dimension (Hausdorff) exists strictly in this limit, at infinity. When the sequence is cut off (quite arbitrarily) and made finite, the incomplete (i.e., euclidean; ref. [17], p. 39) images drawn on paper happen to look like patterns that we see in nature. This coincidence does not mean that natural tree patterns are fractal. The contrary is true: Everything shown to us by nature, and everything done by the fractal algorithm manipulator, supports the view that the real image is euclidean. This is why the natural image can be distinguished by the human eye, because otherwise we would be seeing nothing but blurred images and shades of gray.

Constructal theory is supported by the crisp euclidean images of natural fluid trees that we see. The reason is that the theory starts from the smallest (known, finite) scale, continues as a finite sequence of optimal assemblies, and displays its predictions in two or three dimensions. Along the way, the derived sequence of constructs also explains why the incomplete sequences assumed by the mathematician happen to look like natural patterns.

9. ECONOMICS STRUCTURE IN SPACE

The deterministic power of the constructal principle stretches beyond engineering, physics and biology. The extension to economic structure in space was noted in ref. [1], and is worth emphasizing. Consider the economic activity that covers a given area. The economic activity is the optimization principle, and the structure that covers the area is its result. To see how constructal theory explains the origin of structure in economics and business, consider a stream of goods that proceeds from one point (producer, or factory) to every point of a finite-size territory (consumers). The flow may also proceed in the opposite direction (e.g., grain, carpets woven by individuals). The objective is to minimize the total cost associated with the given stream.

The economies of scale principle tells us that the unit cost is lower when the goods move in the aggregate, i.e., when they are organized into thicker streams. The unit cost is also proportional to the distance traveled. Clearly, the unit cost plays exactly the same role as the local thermal resistance in heat trees, or the local fluid-flow resistance in fluid trees, or the inverse of the travel speed in street trees. The given territory is covered naturally by links of decreasing unit cost, starting from the highest unit cost which is allocated to the smallest area scale (the individual), and continuing with a sequence of intermediaries (distributors) who handle increasingly larger fractions of the given stream of goods.

10. CONSTRUCTAL LAW: THE GENERATING PRINCIPLE FOR GEOMETRIC FORM IN NATURE

In summary, it is possible to deduce from a single statement the shapes of the enormous number of structures of natural systems with internal flows. This 'constructal' law can be stated as follows: for a finite-size open system to persist in time, its configuration must evolve in time in such a way that it provides easier access to the imposed currents that flow through it [1, 2]. This statement has two parts. The first recognizes the natural tendency of imposed currents to construct shapes, i.e. paths of optimal access through constrained open systems. The second part accounts for the evolution (i.e., improvements) of these paths, which occurs in an identifiable direction that can be aligned with time itself.

This formulation of the law refers to a system with imposed steady flow, as in most of the examples reviewed in this article. If the system discharges itself to one point in unsteady fashion, then the geometric minimization of volume-to-point resistance is equivalent to the minimization of the time of discharge, or the maximization of the speed of approach to equilibrium (uniformity, zero flow, death) [16, 18]. If the volume is unbounded, the constructs compound themselves and continue to spread indefinitely, even in three dimensions [19]. Complexity continues to increase in time.

The constructal law defines the concept of necessity, purpose, or optimization. This law is about macroscopic structure. It addresses very old questions that cannot be answered based on known laws: Why should a natural system be optimized? Why should a system be better, or 'more fit' than another (faster, farther, easier, cheaper, etc.)? Why should a system be the 'survivor'? The demonstrated deterministic power of this principle is an invitation to re-examine the classical problems that have evaded determinism in the past.

Acknowledgment

This work was supported by the National Science Foundation.

REFERENCES

[1] Bejan A., Advanced Engineering Thermodynamics, 2nd ed., John Wiley and Sons, New York, 1997.

[2] Bejan A., Int. J. Heat Mass Transfer 40 (1997) 799–816.

[3] Bejan A., Tondeur D., Rev. Gén. Therm. 37 (1998) 165–180.

[4] Tondeur D., Kvaalen E., Ind. Eng. Chem. Res. 26 (1986) 50–56.

[5] Scheidegger A.E., Theoretical Geomorphology, 2nd ed., Springer, Berlin, 1970.

[6] Bejan A., Entropy Generation through Heat and Fluid Flow, John Wiley and Sons, New York, 1982, ch. 4.

[7] Bejan A., Convection Heat Transfer, 2nd edition, Wiley, New York, 1995, ch. 6.

[8] Nelson R.A. Jr., Bejan A., J. Heat Trans.-T. ASME 120 (1998) 357–364

[9] Malkus W.V.R., Proc. Royal Soc. London 225A (1954) 196–212.

[10] Kepler J., The Six-Cornered Snowflake, Oxford University Press, Oxford, UK, 1966 (original in Latin, 1611).

[11] Bejan A., Ikegami Y., Ledezma G.A., Int. J. Heat Mass Tran. 41 (1998) 1945–1954.

[12] Bejan A., Rev. Gén. Therm. 36 (1997) 592–604.

[13] Bejan A., Ledezma G.A., Physica A 255 (1998) 211–217.

[14] Errera M.R., Bejan A., Fractals 6 (1998) 245–261.

[15] Ledezma G.A., Bejan A., Errera M.R., J. Appl. Phys. 82 (1997) 89–100.

[16] Bejan A., in: Ramalingam M.L., Lage J.L., Mei V.C., Chapman J.N. (Eds.), Proceedings of the ASME Advanced Energy Systems Division, AES Vol. 37, ASME, New York, 1997, pp. 257–264.

[17] Mandelbrot B.B., The Fractal Geometry of Nature, Freeman, New York, 1983.

[18] Dan N., Bejan A., J. Appl. Phys. 84 (1998) 3042–3050.

[19] Ledezma G.A., Bejan A., J. Heat Trans.-T. ASME 120 (1998) 977–984.

❏